Journal of Machine Learning Research 23 (2022) 1-53 Submitted 8/20; Revised 6/21; Published 1/22

Model Averaging Is Asymptotically Better Than Model Selection For Prediction

Tri M. Le LE_TM@MERCER.EDU
Department of Science, Mathematics, and Informatics
Mercer University, USA

Bertrand Clarke BCLARKE3@UNL.EDU
Department of Statistics
University of Nebraska-Lincoln, USA 68583-0963

Editor: Sayan Mukherjee

Abstract

We compare the performance of six model average predictors—Mallows' model averaging, stacking, Bayes model averaging, bagging, random forests, and boosting—to the components used to form them. In all six cases we identify conditions under which the model average predictor is consistent for its intended limit and performs as well or better than any of its components asymptotically. This is well known empirically, especially for complex problems, although theoretical results do not seem to have been formally established. We have focused our attention on the regression context since that is where model averaging techniques differ most often from current practice.

Keywords: model averaging, prediction, empirical risk, Mallows, stacking, Bayes, bagging, random forests, boosting

1. The Predictive Perspective

It is a folk theorem that the best predictor for a future outcome is based on the true model, assuming it exists, at least in an asymptotic sense. Results of this nature are often heuristic, see Shmueli (2010) for a discussion. One formal result ensuring this is Rissanen (1984), Theorem 2 and it is clear this result can be extended beyond the model classes he considered. The main exceptions to the principle that the true model gives the best predictions occur when the sample size is insufficient for identifying it. This is often quantified in the concept of variance-bias tradeoff. The mean squared error can increase when the benefit of reducing bias by increasing model complexity is smaller than the corresponding increase in variance. More pragmatically, this commonly occurs when the complexity of the true model e.g., the number of parameters in it, is too high for the sample size. In such cases, a simple but wrong model may give better predictions than a complex but correct model. This situation remains relatively common despite being in the era of big data because 'big data' often means higher per-subject dimension rather than higher sample size relative to the complexity of the data generator (DG). Indeed, even now, to obtain consistency or optimality results, the key adaptation to high model complexity is to allow sample size to increase at a rate that permits the number of parameters to increase as well, albeit slowly, effectively identifying the DG.

©2022 Tri M. Le and Bertrand Clarke.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-874.html.

We take it as given that the best predictors for a future response are derived from the true model, at least asymptotically, assuming it exists. We also tacitly assume that models that are closer to 'true' will give better predictions than those that are further. Then we ask: Suppose the true model is not known, and perhaps cannot even be identified, how do we obtain a good prediction? We give ourselves a list of models to help us identify good predictors assuming that the models were chosen intelligently but do not explore here how that might be done. De facto, we assume that subject matter experts have given us a model list that they think will contain one or more models close to the true model or at least contains models thought to be useful. This is mathematically the same setting as 'prediction with expert advice' as in Cesa-Bianchi and Lugosi (2006).

Our focus here is on model averages for prediction in complex problems and our goal is to show formally that six popular model averages give better predictions than using any of their components would. This is not a new principle—it was observed empirically as early as Galton (1907); see Clemen (1989) and the references therein for a historical perspective. However, given the current interest in prediction, theoretical examination of popular model averages is timely.

In any model average predictor, the two key choices that must be made are the models to average and the way to combine predictions from them. So, suppose we have $\mathcal{F} = \{f_1, \ldots, f_J\}$ as the model list from which to construct a model average. We use the function f_j and the model M_j interchangeably as convenient. Thus $Y_j(x) = f_j(x) + \epsilon_j$ is M_j and we write E_j for the expectation in model M_j where the randomness is in the error terms ϵ_j assumed independent and identically distributed (IID), mean zero, with finite variance σ^2. We assume also that the f_j's are continuously parametrized by real finite dimensional vectors θ_j's and that for each fixed θ_j $f_j(\cdot \mid \theta_j)$ is in a Hilbert space $\mathcal{H}$ with a countable basis and inner product denoted $\langle \cdot, \cdot \rangle$.

Assume we have independent data of the form $\mathcal{D} = \mathcal{D}_n = \{(y_i, x_i) | i = 1, \ldots, n\}$ from $Y(x) = f(x) + \epsilon$ where f is unknown and the x_i's are values of the explanatory variable x. In M_j, $\hat{f}_j(x)$, formed from $\mathcal{D}$, is usually regarded an estimator for $f_j(x)$ and the hope is that at least one of the f_j's is close to f. If we write $\hat{Y}_j$ as the predictor from M_j then consistency of $\hat{f}_j$ (perhaps for f_j) is not the same as assessing how well $\hat{Y}_j$ predicts a generic response Y because $f_j(x)$ is a real number and $Y(x)$ is a random variable. If we proceed sequentially, using $\mathcal{F}$, for $i = 1, \ldots, n+1$ we can use $\mathcal{D}_i$ to define $\hat{Y}(x_{i+1})$ to be a predictor for $Y(x_{i+1})$, the $i+1$ value of the random variable Y at x_{i+1}.

Given the $\hat{Y}_j$'s, we must have a way to combine them to form our overall predictor $\hat{Y}$. Since $\hat{Y}$ is a model average let us denote the weight on the predictor from f_j by α_j. So, we write $\hat{Y}(x_{i+1}) = \sum_{j=1}^{J} \alpha_j \hat{Y}_j(x_{i+1})$. Usually, all $\alpha_j \geq 0$ and they sum to one, i.e., we take a convex combination of the predictors from the M_j's. Even given a fixed $\mathcal{F}$ there are many ways to define the α_j's and to estimate them. (We comment that there are cases where it may be optimal to allow non-negative α_j's to sum to a number different from one—see Clyde and Iversen (2013), Le and Clarke (2016), and Le and Clarke (2017). However, these are cases where a true model cannot be identified or doesn't exist.)

One way to obtain predictors from the f_j's is called the plug-in method. Write $f_j(x_{i+1}) = f_j(x_{i+1} \mid \theta_j)$ and let $\hat{\theta}_j = \hat{\theta}_j(\mathcal{D}_i)$ be an estimator for θ_j. Now, $\hat{Y}_{i+1,j}(x_{i+1}) = f_j(x_{i+1} \mid \hat{\theta}_j) =$

$\hat{f}_j(x_{i+1})$. So, assuming we have a way to form $\hat{\alpha}_j$'s, i.e., $\hat{\alpha}_j = \hat{\alpha}_j(\mathcal{D}_n)$,

$$\hat{Y}(x_{n+1}) = \sum_{j=1}^{J} \hat{\alpha}_j f_j(x_{n+1} \mid \hat{\theta}_j) \tag{1}$$

is the generic form of a model average. Here, we will not explicitly explore the choices of $\mathcal{F}$; our focus is on the $\hat{\alpha}_j$'s.

There are many kinds of theorems that can be proved about model averages such as those in (1). There are three types that we deal with here. The first are consistency theorems for the parameters. That is, results that show how the $\hat{\theta}_j$'s and $\hat{\alpha}_j$'s behave as $n \to \infty$. Usually $\hat{\theta}_j \to \theta_j$ for some θ_j taken as true and $\hat{\alpha}_j \to \alpha_j$ for some limiting α_j, possibly optimal in a sense that allows f to be well approximated by (1). This is fundamentally an estimation perspective, not a prediction perspective.

The second type of theorems are inequalities for the empirical risks of predictors. Often these take the form of oracle inequalities. These are motivated by the definition of the mean squared error and take the generic form

$$\forall \lambda \quad \|\hat{f} - f\|_n \leq C\|f_\lambda - f\|_n + o_P(1) \quad \text{as} \quad n \to \infty, \tag{2}$$

for a predictor $\hat{Y} = \hat{f}$ where $f_\lambda \in \{f_\lambda \in \Lambda\}$, f is true (and may not be f_λ for any λ), $\|\cdot\|_n$ is an empirical norm and $C \in \mathbb{R}^+$. There are numerous variants: The $o(1)$ can often be improved to $\mathcal{O}(1/n)$ when f is a linear model (see Bellec and Tsybakov (2015) for instance) and the case $C = 1$ is called 'sharp'. Also, $o_P(\cdot)$ may be $o(\cdot)$, the empirical norm may be an actual norm, etc. Expressions like (2) are called 'oracle' because it's as if an Oracle knew how to predict Y asymptotically optimally (using $\hat{f}$) without actually knowing the best λ.

To date, oracle inequalities with smaller errors than $o(1)$ seem only to have been established under relatively strong hypotheses such as i) sparsity assumptions on the true model, see Lederer et al. (2019); ii) 'margin' conditions on the set of 'near oracles' to make sure they have small diameter (and hence do not contribute too much complexity), see Lecué (2007) and Yang and Pati (2017), iii) concentration properties on sums of variables, see Lecué and Mitchel (2010), and iv) specific parametric forms of the class of predictors, see Kong and Nan (2014). Indeed, when x is high dimensional and few extra conditions are imposed, $C = (1 + \epsilon)$ and the little-o term amounts to $\mathcal{O}(1/(n\epsilon))$ for $\epsilon > 0$; see Bunea et al. (2004). So, there is a tradeoff beween the empirical risk term and the error term: The larger n is, the smaller ϵ can be.

The third type of theorems are results that show model average predictors are asymptotically better (or no worse) than other predictors. Usually, these latter predictors are taken to be the predictors used to form the average. Indeed, comparisons between a model average and its component predictors are the standard way to verify that the predictive gain from using a model average is worth the cost in interpretability from not simply selecting a single model. Moreover, this is the point of minimizing regret, see Cesa-Bianchi and Lugosi (2006): The 'regret' is the amount by which a predictor could have done better by using one of the models in the average (or 'experts'). It is theorems of this third type that are the main contributions of this paper—although to obtain them we sometimes require theorems of the other two types.

Formally, we want to control $|\hat{Y}(x_{n+1}) - Y(x_{n+1})|$. Unfortunately, there are few theorems, e.g., of type three above, that provide this control. One of the few is due to Raftery and Zheng (2003) for the Bayes model average but even it is on the level of densities not point predictors and does not use a 'true' model. The others that exist are for special cases of linear models, parametric or non-parametric. In the usual notation, we have for parametric linear models ($f_\lambda(x)$ is of the form $X^T\beta$ with $\epsilon \sim N(0, \sigma^2)$) that

$$E(X_{n+1}^T\beta + \epsilon - X_{n+1}^T\hat{\beta})^2 = \sigma^2\left(1 + X_{n+1}^T(X^TX)^{-1}X_{n+1}\right), \tag{3}$$

with similar results for linear estimators in non-parametric function estimation; see Clarke and Clarke (2018) for examples. As predictors, these are, usually, sub-optimal because of model uncertainty and mis-specification.

The intuition behind model average predictors is that they will asymptotically outperform, or at least not underperform, individual model-based predictors, however plausible. These may be found, for example, through model selection. That is, the model average will typically be closer to the DG in an asymptotic and predictive sense than any wrong model and be as good as using the predictor from the true model. Accordingly, we want the model average to be better than any of its components because if we had a model we thought were good, its predictor would already be included in the model average.

Thus, this paper fills a gap in the literature by providing theorems that bound model average predictors in terms of their component predictors i.e., bounding $|\hat{Y}(x_{n+1}) - Y(x_{n+1})|$ in terms of $|\hat{Y}_j(x_{n+1}) - Y(x_{n+1})|$, for six commonly occurring model average predictors. The six model average predictors that we consider are: i) The Mallows' model average (MMA) that minimizes an objective function based on the Mallows model selection principle; ii) the Bayes model averages (BMA) that optimizes a posterior variance; iii) the stacking model average that optimizes a criterion based on cross-validation; iv) random forests (RF's) that tend to stabilize good but unstable predictors; v) bagging more generally; and vi) boosting that optimizes an objective function based on exponential loss, see Friedman et al. (2000), even though it was not originally proposed on that basis.

The structure of this paper is as follows. In Sec. 2, we show theorems of types one, two, and three for the MMA. Being based on linear models, this is a paradigm case in which many of the main features of nonlinear models can be seen easily. In Sec. 3 we study the Bayesian model average predictor under squared error loss—which is what is usually meant by the Bayes model average predictor (BMA). Even though the BMA is not based on linear models, when the models in the average are well-behaved we can show theorems of types one, two, and three and show how some settings with increasing dimension of the explanatory variables can be included. In Sec. 4, we study the stacking model average predictor. Again, we give versions of the three types of results for finite dimensional explanatory variables and discuss the case of increasing dimension of the explanatory variables. In Sec. 5, we turn to the bagging predictor. In this case, there is only one model and the average is formed by bootstrapping so we show versions of the three results but they are of a different character than for MMA, BMA, and stacking. In Sec. 6, we adapt our results from Sec. 5 to the case of random forests (RF) since they are usually bagged trees with a built-in decorrelation procedure. In Sec. 7 we formally deal with boosted regression. Boosting was originally designed for classification so our results here do not address empirical risk but do address consistency and component prediction. As a proxy for empirical risk, however, we are able to observe that the boosted regression predictor is derived from a set of classifiers that achieve the optimal Bayesian risk for classification asymptotically. Finally, in Sec. 8, we provide a more general discussion, in particular on the choice of $\mathcal{F}$. Several technical results are relegated to the Appendices.

2. The Mallows' Model Average (MMA) Predictor

As noted in Sec. 1, there are three qualitatively different sorts of theorems one may seek for a model average predictor assuming the models in the average and the averaging procedure are specified. All three can be seen clearly in the case of linear models. In this case, the consistency result is strong enough to give the empirical risk result fairly easily. The consistency result is also used in the predictive result, but, as will be seen, extra steps are required in the reasoning.

Since our work is an extension of Hansen (2007), we work within his framework. Thus, we write

$$\begin{aligned} Y_i = Y(x_i) &= \mu(x_i) + \epsilon_i \\ &= \sum_{j=1}^{k_m} \theta_j x_{ji} + \sum_{j=k_m+1}^{\infty} \theta_j x_{ji} + \epsilon_i, \end{aligned} \tag{4}$$

in which $E(\epsilon_i \mid x_i) = 0$, $E(\epsilon_i^2 \mid x_i) = \sigma^2$, and $0 \le k_1 \le \cdots \le k_M \le M_n$ where we first think of M_n as a constant and then let it increase slowly with n. In (4) the regression coefficients are the θ_j's and x_{ji} is the j-th component of $x_i = (x_{1i}, \ldots, x_{k_m i})'$. The second term on the right in (4) is the bias b_{mi} and we assume that $E\mu(x_i)^2 < \infty$ and $\mu(x_i)$ converges in mean square for a distribution on x_i. We assume $M = M_n \le n$ is an integer for which $X'_{k_M} X_{k_M}$ is invertible. Also, for $m \le M$, the least squares estimate of Θ_m is $\hat{\Theta}_m = (X'_m X_m)^{-1} X'_m Y_n$ where X_m is the $n \times k_m$ matrix with (i,j) element x_{ji}, $\Theta_m = (\theta_1, \ldots, \theta_{k_m})$, and $Y_n = (Y(x_1), \ldots, Y(x_n))'$.

To define the Mallows' model average (MMA), begin by writing

$$C_n(W) = (Y_n - X_M \hat{\Theta}_M)'(Y_n - X_M \hat{\Theta}_M) + 2\hat{\sigma}^2 k(W), \tag{5}$$

where

$$\begin{aligned} \hat{\Theta}_M &= \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\Theta}_m \\ 0^* \end{pmatrix}, \\ k(W) &= \sum_{m=1}^{M} w_m k_m, \end{aligned} \tag{6}$$

and for given K,

$$\hat{\sigma}_K^2 = \frac{(Y_n - X_K \hat{\Theta}_K)'(Y_n - X_K \hat{\Theta}_K)}{n - K}, \tag{7}$$

where $W = (w_1, \ldots, w_M)$ with $w_1, \ldots, w_M \ge 0$ and $\sum_{m=1}^{M} w_m = 1$. In (6), $\dim \hat{\Theta}_m = k_m$, 0^* means $M - k_m$ repeated 0's, and $k(W)$ is the effective number of parameters. In (7), we use Theorem 2 in Hansen (2007) to see that for $K \to \infty$, and $K/n \to 0$ we get $\hat{\sigma}_K^2 \xrightarrow{P} \sigma^2$. Thus, it is enough to choose $K = k_m$ so that the first term on the right in (4) is a large enough approximation model.

Let $N > 1$ be fixed. Write the M_n-fold Cartesian product of $\{0, 1/N, 2/N, \ldots, 1\}$ as

$$\mathcal{H}_n(N) = \{0, 1/N, 2/N, \ldots, 1\} \times \cdots \times \{0, 1/N, 2/N, \ldots, 1\}$$

and set

$$\hat{W}_N = \arg \min_{W \in \mathcal{H}_n(N)} C_n(W). \tag{8}$$

Now, the MMA for $\mu = (\mu_1, \ldots, \mu_n)'$ is

$$\hat{\mu}_{\hat{W}}(x_1, \ldots, x_n) = X_M \hat{\Theta}_M \tag{9}$$

where $\dim(x_i) = M$ and the M dimensional vector $\hat{W}$, is suppressed in the notation $\hat{\Theta}_M$. The MMA prediction for a new value x_{n+1} of the explanatory variable is

$$\hat{\mu}_{\hat{W}}(x_{n+1}) = x_{n+1} \hat{\Theta}_M. \tag{10}$$

Following Hansen (2007) we evaluate the MMA using squared error, specifically to state a risk consistency result for $\hat{W}_N$. Let $x'_{M,i}$ be the i-th row of X_M. For $\hat{\mu}_i(W)$, the fitted value for x_i using W and the least squares estimate for Θ, write the squared error as

$$\begin{aligned} L_n(W) &= \sum_{i=1}^n \left(\hat{\mu}_i(W) - \mu_i(x_i)\right)^2 \\ &= \sum_{i=1}^n \left(x'_{M,i} \hat{\Theta}_M - \mu_i(x_i)\right)^2 \\ &= \sum_{i=1}^n \left(x'_{M,i} \sum_{m=1}^M w_m \begin{pmatrix} (X'_m X_m)^{-1} X'_m Y_n \\ 0^* \end{pmatrix} - \mu_i(x_i)\right)^2 \\ &= \sum_{i=1}^n \left(\sum_{m=1}^M w_m x'_{M,i} \begin{pmatrix} (X'_m X_m)^{-1} X'_m Y_n \\ 0^* \end{pmatrix} - \mu_i(x_i)\right)^2 . \end{aligned} \tag{11}$$

Now, Theorem 1 in Hansen (2007) gives conditions under which

$$\frac{L_n(\hat{W}_N)/n}{\inf_{W \in \mathcal{H}_n(N)} L_n(W)/n} \xrightarrow{P} 1. \tag{12}$$

Expression (12) is empirical in that $L_n(W)$ it uses $\hat{\Theta}_M$ rather than any limiting values.

Next, we modify the risk consistency result (12) to get a consistency result for $\hat{W}_N$ using the implied true values in $\hat{\Theta}_M$. First rewrite (11) as

$$L_n(W) = \sum_{i=1}^n \left(\sum_{m=1}^M w_m x'_{M,i,m} \hat{\Theta}_m - \mu_i(x_i)\right)^2$$

where $x'_{M,i,m}$ is the first k_m entries of the i-th row of X_M. Taking $M_n = M$ as a constant, we have for each m

$$\hat{\Theta}_m = \left(\frac{1}{n}\sum_{i=1}^{n} x'_i x_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x'_i y_i\right) \tag{13}$$

where x_i is the i-th row of X_m. By the Weak Law of Large Numbers, we get that, for each m, the first term in (13) converges in probability to $EX'_m X_m$ and the second term in (13) converges in probability to $EX'_m Y_n$. So, the continuous mapping theorem applied to $f(X'X, X'Y) = (X'X)^{-1}X'Y$ gives that as $n \to \infty$,

$$\hat{\Theta}_m \overset{P}{\to} (EX'_m X_m)^{-1} E(X'_m Y_n) \equiv \Theta_{\infty,m,M}$$

which is the limit of $\hat{\Theta}_m$ using least squares estimate with design matrix of size M. Under second moment conditions on Y and X we also get

$$\hat{\Theta}_m \overset{L^2}{\to} \Theta_{\infty,m,M}. \tag{14}$$

We use (14) in (13), recalling the bias is $b_{Mi} = \sum_{j=k_M+1}^{\infty} \theta_j x_{ji}$. So

$$\begin{aligned} \frac{1}{n}L_n(W) = \frac{1}{n}\sum_{i=1}^{n} &\left(\sum_{m=1}^{M} w_m x'_{M,i,m}\hat{\Theta}_m - \sum_{j=1}^{M}\theta_j x_{ji}\right)^2 + b_{Mi}^2 \\ &- 2b_{Mi}\left(\sum_{m=1}^{M} w_m x'_{M,i,m}\hat{\Theta}_m - \sum_{j=1}^{M}\theta_j x_{ji}\right) \end{aligned} \tag{15}$$

in which the bias $b_{Mi} \overset{L^2}{\to} 0$ as $M \to \infty$. This means that the last two terms in (15) go to zero in L^2 and hence in probability and distribution. Now,

$$\frac{1}{n}L_n(W) \overset{P}{\to} E\left(\sum_{m=1}^{M} w_m x'_{M,i,m}\Theta_{\infty,m,M} - \sum_{j=1}^{M}\theta_j x_{ji}\right)^2 \equiv L_{\infty,M}(W)$$

where $L_{\infty,M}(W)$ is the limit, pointwise in W. Now, by Markov's inequality, fourth moment conditions on the terms in $L_n(W)$ give that

$$\frac{1}{n}L_n(W) \overset{P}{\to} L_{\infty,M}(W) \tag{16}$$

uniformly on compact sets of W in $[0,1]^M$.

We have the following consistency result for $\hat{W}_N$.

Theorem 1 *: Assume the hypotheses of Theorems 1 and 2 in Hansen (2007), $E(Y^4), E(\|X\|^4) < \infty$ and that $L_{\infty,M}(\cdot)$ has a unique minimum at $W_{opt,M}$. Assume also that as $M \to \infty$, $w_{opt,M,j} \to w_{opt,\infty,j}$ i.e. as $M \to \infty$ and $n \to \infty$ the entries of $W_{opt,M}$ converge to limits as well. Then,*

$$\hat{W}_N \overset{P}{\to} W_{opt,M} \tag{17}$$

and there is a limiting vector W_∞ so that for each M, $W_{opt,M} \to W_{\infty,M}$, the first M entries of W_∞, as $n, N \to \infty$ at appropriate rates.

Proof : For each M, the uniformity in (16) gives

$$\min_W \frac{1}{n} L_n(W) \overset{P}{\to} \min_W L_{\infty,M}(W) \tag{18}$$

If the $\min_W$ is over $W \in \mathcal{H}_n(N)$ then (18) with (12) gives

$$\frac{L_n(\hat{W}_N)}{n} \overset{P}{\to} \min_{W \in \mathcal{H}_n(N)} L_{\infty,M}(W) \tag{19}$$

Since $L_{\infty,M}(\cdot)$ has a unique minimum, the Newey-McFadden Theorem (see Theorem 2.1 in Newey and McFadden (2012)) gives (17). ■

This is the appropriate consistency result for MMA. The weights converge to limits as do the parameters in the functions being averaged.

The second kind of result we want is the limiting behavior of the empirical risk. This follows easily from (12) because it shows that the MMA asymptotically achieves the minimal empirical squared error. Indeed, we can see that the MMA is better than any individual model simply considering the weight vector W that puts $w_j = 1$ and $w_{j'} = 0$ for $j' \neq j$. For this vector, the MMA is asymptotically better than just using the j-th model alone. The same reasoning applies if $L_n(W)$ is replaced by its less empirical form

$$L_n(\hat{W}_N) = \sum_{i=1}^{n} \left(\sum_{m=1}^{M} \hat{w}_m x'_{M,i,m} \Theta_{\infty,m,M} - \mu_i(x_i) \right)^2 \tag{20}$$

that has limit as in (16). It is easily seen that (20) and its empirical form can be expressed as an oracle inequality such as (2).

Next we turn to the third kind of result we want—this is the most important of the three results because it addresses predictive performance directly. Our result asymptotically comparing the squared error for individual predictions from individual models with predictions from the MMA is the following.

Theorem 2 *Under the hypotheses of Prop. 1, if we let $M = M_n \to \infty$ with $n \to \infty$, slowly, and $K \to \infty$ and $N \to \infty$ but slow enough that the convergences for $m = 1, \ldots, M$ are uniformly good then*

$$\begin{aligned} \lim_{n\to\infty} E\Big(Y(x_{n+1}) - x_{n+1}(X'_M X_M)^{-1} X'_M Y_n\Big)^2 & \\ - E\left(Y(x_{n+1}) - \sum_{m=1}^{M} \hat{w}_m x_{n+1}(X'_m X_m)^{-1} X'_m Y_n\right)^2 &\geq 0. \end{aligned} \tag{21}$$

Proof By recalling $Y = \mu + \epsilon$ and bias $\overset{L^2}{\to} 0$, (21) becomes

$$\begin{aligned} \lim_{n\to\infty} E\Big(\mu(x_{n+1}) - x_{n+1}(X'_M X_M)^{-1} X'_M Y_n\Big)^2 & \\ - E\left(\mu(x_{n+1}) - \sum_{m=1}^{M} \hat{w}_m x_{n+1}(X'_m X_m)^{-1} X'_m Y_n\right)^2 &\geq 0. \end{aligned} \tag{22}$$

(Recall we are setting $k_M = M$ for convenience.) Simplifying (22) by using $\mu(x_{n+1}) = \sum_{j=1}^{M} \theta_j x_{j,n+1} + \sum_{j=M+1}^{\infty} \theta_j x_{j,n+1}$ gives that it is equivalent to

$$\begin{aligned}\lim_{n\to\infty} E & \left(\sum_{j=1}^{M} \theta_j x_{j,n+1} - x_{n+1}\hat{\Theta}_M\right)^2 \\ & \qquad - E\left(\sum_{j=1}^{M} \theta_j x_{j,n+1} - \sum_{m=1}^{M} \hat{w}_m x_{n+1}\hat{\Theta}_m\right)^2 \geq 0.\end{aligned} \tag{23}$$

Examining (23) for fixed $k_M = M_n = M$, $\hat{w}_m \to w_{\infty,m,opt}$, and $\hat{\Theta}_m \to \Theta_{\infty,m,M}$, we have the second term in (23) goes to

$$E\left(\sum_{j=1}^{M} \theta_j x_{j,n+1} - \sum_{m=1}^{M} w_{\infty,m,opt} x_{n+1}\Theta_{\infty,m,M}\right)^2 = \inf_W L_{\infty,M}(W). \tag{24}$$

On the other hand, the first term in (23) goes to

$$\begin{aligned} & E\left(\sum_{j=1}^{M} \theta_j x_{j,n+1} - x_{n+1}\Theta_{\infty,M,M}\right)^2 = L_{\infty,M}((0,\ldots,0,1)) \\ & \geq \inf_W L_{\infty,M}(W).\end{aligned} \tag{25}$$

Taken together, (24) and (25) give (23). ■

Theorem 2 shows that the MMA will never asymptotically underperform any of its component models. This is the key property we want any model average predictor to satisfy.

3. Bayesian Model Averaging

Bayesian model averaging BMA is another model averaging technique that takes model uncertainty into account by using the posterior weights of models. BMA was first developed in Leamer (1978); see Geisser (1993), Draper (1995), and Raftery et al. (1996), among others. Skouras and Dawid (1998) (Theorem 4) established the efficiency of BMA and Raftery and Zheng (2003) (Theorems 2 and 4) established other asymptotic optimality properties of BMA under logarithmic scoring rules. These results show that taking model uncertainty into account improves prediction, see Clyde and George (2004). Taken together, these show the efficacy of BMA for prediction but do not actually show that the BMA predictor is better than the predictors of any of it components asymptotically. Here, we establish this formally.

The central idea of BMA is as follows. Suppose we have J models $f_j(x \mid \theta_j)$, $j = 1, \ldots, J$ and $Y(x) = f(x \mid \theta_j) + \epsilon$ with density $p_j(y \mid x, \theta_j)$ but that j is unknown. (Later we give remarks on cases where $Y(x) = f_T(x) + \epsilon$ i.e., where no f_j is true.) Equip each θ_j with a prior $w(\theta_j \mid M_j)$, where M_j indicates the j^{th} model f_j, and let $W(M_j)$ be the across-models prior for the M_j's. When convenient we write $p(y_i \mid \theta_j, M_j) = p(y_i \mid \theta_j)$ or use other simplified notation provided what is meant is clearly indicated. Given data $\mathcal{D}_n$, the predictive distribution of the future value Y_{n+1} is

$$p(y_{n+1} \mid \mathcal{D}_n) = \sum_{j=1}^{J} p(y_{n+1} \mid M_j, \mathcal{D}_n) W(M_j \mid \mathcal{D}_n). \tag{26}$$

This is an average of the conditional predictive distributions $p(y_{n+1} \mid M_j, \mathcal{D}_n)$ weighted by the posterior probability of M_j, $W(M_j \mid \mathcal{D}_n)$, in which

$$p(y_{n+1} \mid M_j, \mathcal{D}_n) = \int p(y_{n+1} \mid \theta_j, M_j) w(\theta_j \mid M_j, \mathcal{D}_n) d\theta_j,$$
$$W(M_j \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid M_j) W(M_j)}{\sum_{j=1}^{J} p(\mathcal{D}_n \mid M_j) W(M_j)},$$

and the marginal likelihood under model M_j is

$$p_j(y \mid x, M_j) = \int p_j(y \mid \theta_j, x, M_j) w_j(\theta_j \mid M_j) d\theta_j.$$

We also write $w(\theta_j \mid M_j, \mathcal{D}_n) = w_j(\theta_j \mid \mathcal{D}_n)$ for the posterior for θ_j given the data $\mathcal{D}_n$. The 'pure' BMA (26) is defined literally as an average of probability models and hence is a probability model itself. Here, we only use one of the predictors that can be derived from this model average. Specifically, the BMA predictor under squared error loss for Y_{n+1} is

$$\hat{Y}_{BMA}(x_{n+1}) = E(Y_{n+1}(x_{n+1}) \mid \mathcal{D}_n) = \sum_{j=1}^{J} W(M_j \mid \mathcal{D}_n) E(Y_{n+1}(x_{n+1}) \mid M_j, \mathcal{D}_n) \tag{27}$$

in which

$$E(Y_{n+1}(x_{n+1}) \mid M_j, \mathcal{D}_n) = \int y_{n+1}(x_{n+1}) p(y_{n+1}(x_{n+1}) \mid M_j, \mathcal{D}_n) dy_{n+1}.$$

Being fully Bayesian, the consistency properties of the parameters in a BMA predictor can be stated and established using standard techniques. There are two kinds of parameters in (27): The parameters θ_j in each of the M_j's and the M_j's themselves. That is, the posterior weight of a model M_j is also de facto a parameter. Our first result is on consistency and is the following.

Theorem 3 *: Assume the hypotheses of Lemma 1, Lemma 2, and Lemma 4 hold for $p_j(y \mid \theta_j, x, M_j)$ for $j = 1, \ldots, J$. Also assume the J parametric families are jointly soundly parametrized, i.e.,*

$$\min_{j, j', \theta_j, \theta_{j'}, x} D(P_{j,\theta_j,x} \| P_{j',\theta_{j'},x}) > c \tag{28}$$

for some $c > 0$ and that the limit of $(1/n) \sum_{i=1}^{n} I_i(\theta^*_{j^*} \mid x_i)$ exists as $n \to \infty$. Finally, we assume that the parameter space and the range of x-values is the same for all j and both are compact subsets of d-dimensional real space.

Then, i) for all $j = 1, \ldots, J$ we have that there are values $\hat{\theta}^*_{j^*}$ so that $\hat{\theta}_j \to \theta^*_{j^*}$ when M_{j^*} is taken as true and ii) the posterior weights $W(M_j \mid \mathcal{D}) \to 0, 1$ under M_{j^*} when $j^* = j$ and $j^* \neq j$, respectively.

Remark: The condition (28) is stronger than is required for convergence alone. We impose it because shortly we will be approximating $\hat{Y}_{BMA}$ and want the approximation to be well-defined.

Proof : First, we verify convergence of the $\hat{\theta}_j$'s—though not necessarily consistency to the true values. To begin, observe that Lemma 2 ensures the posterior mean of any θ_j, when j is taken as true, exists and converges to its correct value θ_j^*. Also, for any j taken as true, the hypotheses of Lemma 4 ensure there exist values θ_{j*}^*, for $j^* \neq j$, to which the posterior mean $E(\theta_j \mid M_j, \mathcal{D}_n)$ converges for any given M_{j^*}. The value $\theta_{j^*}^*$ need not be equal to the limit of the posterior mean $E(\theta_j \mid M_j, \mathcal{D}_n)$ when the data comes from M_j. Thus, each $\hat{\theta}_j$ converges even if it is only consistent under M_j.

Second, we show that the posterior model probabilities converge to one or zero according to whether the model is true or not. Since we are not in the wrong-model case, write j^* for the index of the true parametric family. Now, for any $j \neq j^*$ we have

$$W(M_j \mid \mathcal{D}_n) = \frac{p_j(\mathcal{D}_n \mid M_j) W(M_j)}{\sum_{j=1}^J p_j(\mathcal{D}_n \mid M_j) W(M_j)}, \tag{29}$$

where $p_j(\mathcal{D}_n \mid M_j) = \int w_j(\theta_j) p_j(y^n \mid \theta_j, x^n, M_j) d\theta_j$. We want $W(M_j \mid \mathcal{D}_n) \to 0$ for $j \neq j^*$, in $p(y \mid \theta_{j*}^*, x, M_{j*})$ since this will also give $W(M_{j*} \mid \mathcal{D}_n) \to 1$. Rewriting (29) gives

$$W(M_j \mid \mathcal{D}_n) = \frac{\frac{\int w_j(\theta_j) p_j(y^n | \theta_j, x^n, M_j) d\theta_j W(M_j)}{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n | \theta_{j^*}, x^n, M_{j^*}) d\theta_{j^*} W(M_{j^*})}}{1 + \sum_{j=1, j \neq j^*}^J \frac{\int w_j(\theta_j) p_j(y^n | \theta_j, x^n, M_j) d\theta_j W(M_j)}{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n | \theta_{j^*}, x^n, M_{j^*}) d\theta_{j^*} W(M_{j^*})}}. \tag{30}$$

It is seen that the numerator and denominator in (30) are based on ratios of the same form, namely

$$\frac{\int w_j(\theta_j) p_j(y^n \mid \theta_j, x^n, M_j) d\theta_j W(M_j)}{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n \mid \theta_{j^*}, x^n, M_{j^*}) d\theta_{j^*} W(M_{j^*})}. \tag{31}$$

So, it is enough to show that ratios like (31) go to zero in probability for $j \neq j'$. To this end write (31) as

$$\begin{aligned} \frac{W(M_j)}{W(M_{j*})} \int w_j(\theta_j) \exp \Bigg[-\Bigg(& \ln \frac{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n \mid \theta_{j^*}, x^n, M_{j^*}) \mathrm{d}\theta_{j*}}{p_{j*}(y^n \mid \hat{\theta}_{j*}, x^n, M_{j*})} \\ & + \ln \frac{p_{j*}(y^n \mid \hat{\theta}_{j*}, x^n, M_{j*})}{p_{j*}(y^n \mid \theta_{j*}^*, x^n, M_{j*})} + \ln \frac{p_{j*}(y^n \mid \theta_{j*}^*, x^n, M_{j*})}{p_j(y^n \mid \theta_j, x^n, M_j)} \Bigg) \Bigg] \mathrm{d}\theta_j, \end{aligned} \tag{32}$$

where $\hat{\theta}_{j*}$ is the MLE under M_{j*}.

The third term in parentheses in (32) dominates the other two. Indeed, we see that it equals

$$n \left(\frac{1}{n} \sum_{i=1}^n \ln \frac{p_{j*}(y_i \mid \theta_{j*}^*, x_i, M_{j*})}{p_j(y_i \mid \theta_j, x_i, M_j)} \right). \tag{33}$$

The expectation of (33) (with respect to the density $p_{j*}(y_i \mid \theta^*_{j*}, x_i, M_{j*})$) gives a relative entropy that, by the soundness condition, is strictly positive. By the law of large numbers, this relative entropy is the limit of the term in parentheses. So, (33) is $\mathcal{O}_P(n)$.

The first term in parentheses in (32) is of order $\mathcal{O}_P(\ln n)$. To see this note that Lemma 1 ensures the INID MLE converges to its limit for the sequence of x_i's being used. By a standard Laplace's method argument, see Clarke and Barron (1988) Appendix A or DeBruijn (1958), we can show that

$$\Bigg| \ln \frac{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n \mid \theta_{j^*}, x^n, M_{j^*}) \mathrm{d}\theta_{j*}}{p_{j*}(y^n \mid \hat{\theta}_{j*}, x^n, M_{j*})} \\ - \frac{d_j}{2} \ln \frac{2\pi}{n} - \left(\det \left(\frac{1}{n} \sum_{i=1}^{n} \hat{I}_i(\hat{\theta}_{j^*} \mid x_i) \right)^{1/2} \right) - \ln w_{j^*}(\hat{\theta}_{j^*}) \Bigg| \to 0 \tag{34}$$

in $P_{\theta^*_{j*}}$-probability. Standard arguments give that the estimates can be replaced by their limits so we have

$$\ln \frac{\int w_{j^*}(\theta_{j^*}) p_{j^*}(y^n \mid \theta_{j^*}, x^n, M_{j^*}) \mathrm{d}\theta_{j*}}{p_{j*}(y^n \mid \hat{\theta}_{j*}, x^n, M_{j*})} \\ \approx \frac{d_{j^*}}{2} \ln \frac{n}{2\pi} + \left(\det \left(\frac{1}{n} \sum_{i=1}^{n} I_i(\theta^*_{j^*} \mid x_i) \right)^{1/2} \right) + \ln w_{j^*}(\theta^*_{j^*}). \tag{35}$$

So, the first term is $\mathcal{O}(\ln n)$ and hence dominated in probability by $\mathcal{O}_P(n)$.

The middle term in (32) has limiting behavior from Wilks' theorem and so is of order strictly smaller than any $\mathcal{O}_P(a_n)$ for $a_n \to \infty$. In particular, we can choose $a_n = \ln n$. To see this, note the assumptions of Lemma 1 give

$$2 \ln \frac{p_{j^*}(y^n \mid \hat{\theta}_{j^*}, x^n, M_{j^*})}{p_{j^*}(y^n \mid \theta^*_{j^*}, x^n, M_{j^*})} = 2(\ell_n(\hat{\theta}_{j^*} \mid \mathcal{D}_n) - \ell_n(\theta^*_{j^*} \mid \mathcal{D}_n)),$$

where ℓ_n is the log likelihood function. Using Taylor's Theorem to expand $\ell_n(\theta_{j^*} \mid \mathcal{D}_n)$ about $\hat{\theta}_{j^*}$ gives

$$\ell_n(\theta_{j^*} \mid \mathcal{D}_n) - \ell_n(\hat{\theta}_{j^*} \mid \mathcal{D}_n) = \dot{\ell}_n(\hat{\theta}_{j^*} \mid \mathcal{D}_n)(\theta_{j^*} - \hat{\theta}_{j^*}) - n(\theta_{j^*} - \hat{\theta}_{j^*})^T A_n(\theta_{j^*} \mid \mathcal{D}_n)(\theta_{j^*} - \hat{\theta}_{j^*}).$$

It is seen that $\dot{\ell}_n(\hat{\theta}_{j^*} \mid \mathcal{D}_n) = 0$ and from the proof of Lemma 2

$$A_n(\theta_{j^*} \mid \mathcal{D}_n) = -\frac{1}{n} \int_0^1 \int_0^1 v \ddot{\ell}_n(\hat{\theta}_{j^*} + uv(\theta_{j^*} - \hat{\theta}_{j^*})) du dv \\ \approx \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} I_i(\theta_{j^*} \mid x_i).$$

Now, by mild abuse of notation, we have

$$2 \ln \frac{p_{j^*}(y^n \mid \hat{\theta}_{j^*}, x^n, M_{j^*})}{p_{j^*}(y^n \mid \theta^*_{j^*}, x^n, M_{j^*})} = 2n(\theta_{j^*} - \hat{\theta}_{j^*})^T A_n(\theta_{j^*} \mid \mathcal{D}_n)(\theta_{j^*} - \hat{\theta}_{j^*}) \\ \approx n(\theta_{j^*} - \hat{\theta}_{j^*})^T \left[\frac{1}{n} \sum_{i=1}^{n} I_i(\theta_{j^*} \mid x_i) \right] (\theta_{j^*} - \hat{\theta}_{j^*}).$$

Lemma 1 gives that $\sqrt{n}(\hat{\theta}_{j^*} - \theta_{j^*}) \approx N\left(0, [1/n \sum_{i=1}^{n} I_i(\theta_{j^*} \mid x_i)]^{-1}\right)$. Using this in the last expression gives that asymptotically

$$2 \ln \frac{p_{j^*}(y^n \mid \hat{\theta}_{j^*}, x^n, M_{j^*})}{p_{j^*}(y^n \mid \theta^*_{j^*}, x^n, M_{j^*})} \xrightarrow{\mathcal{L}} \chi^2_{d_{j^*}} \tag{36}$$

which is smaller than any $\mathcal{O}_P(a_n)$ with $a_n \to \infty$; e.g., $a_n = \log\log n$.

Our analysis of (32) shows it is $\exp(-\mathcal{O}_P(n) - \mathcal{O}_P(\ln n) - \mathcal{O}_P(a_n))$, where α is the limiting factor on the leading term (n in the exponent) and is the limit of the average in (33). Clearly, (30) is essentially of the form $(31)_j/(1+ \sum_{j^* \neq j}(31))$. So, we can multiply the numerator and the denominator by $e^{\alpha n}$. Consequently, (30) is of the form

$$\frac{1 + e^{\alpha n} e^{-\mathcal{O}_P(\ln n) - \mathcal{O}_P(a_n)}}{e^{\alpha n} + \sum_{j^* \neq j}(1 + e^{\alpha n} e^{-\mathcal{O}_P(\ln n) - \mathcal{O}_P(a_n)})}$$

where the numerator is for j. Thus, we see that for $j \neq j^*$, $W(M_j \mid \mathcal{D}_n) \to 0$ in probability as $n \to \infty$ and consequently $W(M_{j^*} \mid \mathcal{D}_n) \to 1$. ■

Next, we compare the empirical risk of BMA with the empirical risk of its component models. As discussed in Secs. 1 and 2, this is the second sort of result that is useful with model averages. So, suppose j^* indexes the true model. Then, the empirical risk of M_{j^*} is

$$\begin{aligned}
L_n(M_{j^*}) &= \sum_{i=1}^{n} (y_i - E_{j^*}(Y(x_i) \mid \mathcal{D}_n))^2 \\
&= \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n))^2 \\
&+ \sum_{i=1}^{n} (\sum_{j=1}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n) - E_{j^*}(Y_i \mid \mathcal{D}))^2 \\
&+ 2 \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n))(\sum_{j=1}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n) - E_{j^*}(Y_i \mid \mathcal{D})) \qquad (37) \\
&= L(BMA) + \sum_{i=1}^{n} (\sum_{j \neq j^*}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n) + (W(j^* \mid \mathcal{D}) - 1) E_{j^*}(Y(x_i) \mid \mathcal{D}_n))^2 \\
&+ 2 \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n)) \\
&\times (\sum_{j \neq j^*}^{J} W(j \mid \mathcal{D}) E_j(Y(x_i) \mid \mathcal{D}_n) + (W(j^* \mid \mathcal{D}) - 1) E_{j^*}(Y(x_i) \mid \mathcal{D}_n).
\end{aligned}$$

To obtain the desired bounds on empirical risks, we show that the last two terms on the right in (37) go to zero at an exponential rate in probability. Fundamentally, this is a property of Bayes consistency on a discrete space namely the model indices $j = 1, \ldots, J$. Since Bayes consistency is exponentially fast we have the following.

Theorem 4 *: Assume that for each $\theta_j \in M_j$ for $j = 1, \ldots, J$ we have that*

$$E_j \sup_{|\theta_j' - \theta_j| \leq \delta} |\frac{\partial^2}{\partial\theta_{j,k}\partial\theta_{\jmath,\ell}} \log p_j(X \mid \theta')|^2 < B$$

for k, ℓ ranging over $1, \ldots, \dim(\theta_j)$ where $B > 0$. Also, assume that the Renyi relative entropy of order $1 + \lambda$

$$\frac{1}{\lambda} \log \int p_j(x \mid \theta_j) \left(\frac{p_j(x \mid \theta_j)}{p_j(x \mid \theta_j'} \right)^{\lambda} \mathrm{d}x$$

is uniformly bounded over θ_j and θ_j' for all j. Then, for $j = 1, \ldots, J$, when M_j is true, as $n \to \infty$, we have

$$|L_n(BMA) - L_n(M_{j^*})| = o_P\left(e^{-\alpha n}\right), \tag{38}$$

for some $\alpha > 0$.

Proof : It is enough to examine the posterior probabilities in last two terms on the right of (37). Observe that the hypotheses of Theorem 4 include the hypotheses of Proposition 1 in Clarke (1999). Thus, this Proposition applied in the independent not identical case for the mixed continuous and discrete parameter space $(\Theta_1 \cup \ldots \cup \Theta_J) \times \{1, \ldots, J\}$ gives that the model weights converge where they should exponentially fast in probability. That is, there is a $\alpha' > 0$ so that in (37), $P_j(W(j \mid \mathcal{D}) \geq e^{-\alpha n}) \leq e^{-\alpha' n}$ for $j \neq j^*$ and $P_{j^*}(W(j^* \mid \mathcal{D}) \geq e^{-\alpha n}) \leq e^{-\alpha' n}$. Using this with bounds on the posterior means gives that each term on the right in (37) is $\mathcal{O}_p(e^{-\alpha' n})$. Since there are n terms (the sums over $i = 1, \ldots, n$), it is enough to choose a value $\alpha \in (0, \alpha')$ to satisfy (38). ■

Expression (38) is evidence that BMA provides no worse prediction on average, under squared error loss, than using the predictor from the best model in the BMA and it is easy to modify the proof to see that $L(BMA) \leq L(M_{j^*}) + \mathcal{O}(e^{-\alpha n})$ and $L(M_{j^*}) < L(M_j) + C(j, j^*) + \mathcal{O}(e^{-\alpha n})$ for $j \neq j^*$ where $C(j, j^*) > 0$ depends on how close the j-th and j^*-th parametric families are.

However, Theorem 4 is in cumulative error over n predictions, not the error of an individual prediction. To address individual predictions, we state and prove a result in the same spirit as Theorem 2 but for $\mathcal{M}$-complete and $\mathcal{M}$-closed DG's.

We begin by noting that expression (27) has a convenient approximation. If M_j is the true model so that $Y(x) = f_j(x \mid \theta_j) + \epsilon$ we have

$$E(Y_{n+1}(x_{n+1}) \mid M_j, \mathcal{D}_n) = \int y_{n+1}(x_{n+1}) m_j(y_{n+1}(x_{n+1}) \mid \mathcal{D}_n) dy_{n+1}$$

and we are led to consider the approximation to the BMA given by

$$\hat{Y}_{BMA,app}(x) = \sum_{j=1}^{J} W(M_j \mid \mathcal{D}_n) f_j(x \mid E(\theta_j \mid \mathcal{D}_n)), \tag{39}$$

where $E(\theta_j \mid \mathcal{D}_n)$ is a the posterior mean estimator for $\theta_{j,T}$ when model M_j is true and for $\theta_{j^*}^*$ when M_{j^*} is true. As a first step to showing Theorem 5 we verify that (39) and (27) are asymptotically close.

Proposition 1 *: Assume the hypotheses of Theorem 3. Then,*

$$\mid \hat{Y}_{BMA}(x) - \hat{Y}_{BMA,app}(x) \mid \overset{P}{\to} 0 \textit{ pointwise in } x, \tag{40}$$

in M_{j^*}, whichever of the J models is the overall true model.

Remarks: First, for linear models $\hat{Y}_{BMA,app}(x) = \hat{Y}_{BMA}(x)$. So, the Proposition is immediate. Second, we conjecture that this proposition holds not just in the INID setting but also in the wrong-model INID setting i.e., when the true model is not in any of the M_j's. That is, there is some INID M_T distinct from all the INID M_j's in which convergence must be assessed. In the notation of Berk (1966) (for the IID case), if we only use one M_j when M_T is true then $\theta_j^* = \arg\min_{\theta_j} \eta_j(\theta_j)$ and the minimum is denoted $\eta_j^* = \eta^*(\theta_j^*)$. In White (1982), we have $\eta_j(\theta) = E_T \ln(p_T(y)/p_j(y \mid \theta_j))$ where T indicates the true density. So, following White (1982) in the IID case, $\theta_j^* = \arg\min D(p_T \| p_{\theta_j})$. (This must be averaged over x but we have omitted this for ease of exposition.) So, it seems intuitive that the posterior model weight corresponding to

$$(j^*, \theta^*) = \arg\min_{j,\theta_j} D(P_T \| P_{j,\theta_j})$$

will go to one and the posterior model weights for other M_j's will go to zero. We believe this will follow from Lemmas 3 and 4 that generalize Berk (1966) and White (1982) to the wrong-model INID settings. Note that in the present case we are using multiple models M_j. So, there are 'wrong' models. However, we are assuming that one of the M_j's is correct.

Proof : For any j taken as true, the hypotheses of Lemma 4 ensure there exist values $\theta_{j^*}^*$, for $j^* \neq j$, to which the posterior mean $E(\theta_j \mid M_j, \mathcal{D}_n)$ converges for any given M_{j^*}. The value $\theta_{j^*}^*$ need not be equal to the posterior mean of $E(\theta_j \mid M_j, \mathcal{D}_n)$ when the data comes from M_j. Lemma 2 ensures the posterior mean of any θ_j, when j is taken as true, exists and converges to its correct value θ_j^*. This ensures that $\hat{Y}_{BMA,app}$ is well-defined regardless of which M_j is true.

By using these facts about the posterior means and then using Theorem 3 to identify the limits of the posterior model weights we have when j^* indexes the true model class that

$$\begin{aligned}
&\mid \hat{Y}_{BMA}(x_{n+1}) - \hat{Y}_{BMA,app}(x_{n+1}) \mid \\
&= \left| \sum_{j=1}^{J} W(M_j \mid \mathcal{D}_n) \left[E(Y_{n+1}(x_{n+1}) \mid M_j, \mathcal{D}_n) - f_j(x_{n+1} \mid E(\theta_j \mid M_j, \mathcal{D}_n)) \right] \right| \\
&\approx \left| E(Y_{n+1}(x_{n+1}) \mid M_{j^*}, \mathcal{D}_n) - f_{j^*}(x_{n+1} \mid E(\theta_{j^*} \mid M_{j^*}, \mathcal{D}_n)) \right|.
\end{aligned}$$

Both terms in the last expression have the same limit, namely $f_{j^*}(x_{n+1} \mid \theta_{j^*}^*)$. So, the statement of Prop. 1 follows. ■

Our analog to Theorem 2 for BMA is the following. It is different in character from Theorem 2 (and Theorem 8 in Sec. 4) because posterior model weights generally only converge to zero or one as seen in Theorem 3. That is, they are doing model selection more than model approximation. Again, this result is for the INID but not wrong-model setting.

Theorem 5 *Assume the hypotheses of Prop. 1 and that for some fixed j^* $Y_T(x) = Y(x) = f_{j^*}(x \mid \theta^*_{j^*}) + \epsilon$ where ϵ has mean zero and variance σ^2. Then, provided each $\hat{\theta}_j = E(\theta_j \mid M_j, \mathcal{D}_n)$ is consistent for some θ^*_j under model f_{j^*} as $n \to \infty$ we have*

$$\limsup_{n\to\infty} \left(\int E_Y(Y(x) - f_j(x \mid \hat{\theta}_j))^2 \mathrm{d}x - \int E_Y(Y(x) - \hat{Y}_{BMA}(x))^2 \mathrm{d}x \right) \geq 0.$$

Proof We have

$$\begin{aligned} E_Y(Y(x) - \hat{Y}_{BMA}(x))^2 &= E_Y(Y(x) - \hat{Y}_{BMA}(x) \pm EY(x))^2 \\ &= E_Y(Y(x) - EY(x))^2 + E_Y(EY(x) - \hat{Y}_{BMA}(x))^2 \\ &\quad + 2E_Y \left[(Y(x) - EY(x))(EY(x) - \hat{Y}_{BMA}(x)) \right]. \end{aligned}$$

The last term is zero because the first factor in this term is for x_{n+1} and the other factor in this term is for $x_i, i \leq n$. So, the factors are independent, the expectation factors, and the first expectation is zero. We repeat this argument by adding and subtracting $\hat{Y}_{BMA,app}$ to get

$$\begin{aligned} E_Y(Y(x) - \hat{Y}_{BMA}(x))^2 &= \sigma^2 + E_Y(EY(x) - \hat{Y}_{BMA}(x) \pm \hat{Y}_{BMA,app}(x))^2 \\ &= \sigma^2 + E_Y(EY(x) - \hat{Y}_{BMA,app}(x))^2 + E_Y(\hat{Y}_{BMA}(x) - \hat{Y}_{BMA,app}(x))^2 \\ &\quad + 2E_Y \left[(EY(x) - \hat{Y}_{BMA,app}(x))(\hat{Y}_{BMA}(x) - \hat{Y}_{BMA,app}(x)) \right] \end{aligned}$$

By the compactness assumptions of Prop. 1 and the bounded moment conditions assumed earlier, we can use the statement of Prop. 1 i.e., $\hat{Y}_{BMA}(x) - \hat{Y}_{BMA,app}(x) \overset{P}{\to} 0$ for each x, to see that the third and fourth terms goes to 0 pointwise in x. Hence, asymptotically,

$$\begin{aligned} E_Y(Y(x) - \hat{Y}_{BMA}(x))^2 &= \sigma^2 + E_Y(EY(x) - \hat{Y}_{BMA,app}(x))^2 + o_P(1) \\ &= \sigma^2 + E_Y\Big(f_{j^*}(x \mid \theta^*_{j^*}) - \sum_{j=1}^{J} W(M_j \mid \mathcal{D}_n) f_j(x \mid E(\theta_j \mid M_j, \mathcal{D}_n)) \Big)^2 + o_P(1) \\ &\approx \sigma^2 + E_Y\Big(f_{j^*}(x \mid \theta^*_{j^*}) - f_{j^*}(x \mid E(\theta_{j^*} \mid M_{j^*}, \mathcal{D}_n)) \Big)^2 \\ &\approx \sigma^2 + \Big(f_{j^*}(x \mid \theta^*_{j^*}) - f_{j^*}(x \mid \theta^*_{j^*}) \Big)^2 \approx \sigma^2, \end{aligned} \tag{41}$$

where $\theta^*_{j^*}$ is the correct value θ_{j^*} from the best $f_{j^*}(x \mid \theta_{j^*})$.

For the other side of the inequality in Theorem 5 we fix j, use similar arguments as before and get

$$\begin{aligned} E_Y(Y(x) - f_j(x \mid \hat{\theta}_j))^2 &= E_Y(Y(x) - f_j(x \mid \hat{\theta}_j) \pm EY(x))^2 \\ &= E_Y(Y(x) - EY(x))^2 + E_Y(EY(x) - f_j(x \mid \hat{\theta}_j))^2 \\ &\quad +2E_Y \left[(Y(x) - EY(x))(EY(x) - f_j(x \mid \hat{\theta}_j)) \right] \\ &= \sigma^2 + E_Y\Big(f_{j^*}(x \mid \theta^*_{j^*}) - f_j(x \mid E(\theta_j \mid M_T, \mathcal{D}_n)) \Big)^2 \\ &\approx \sigma^2 + \Big(f_{j^*}(x \mid \theta^*_{j^*}) - f_j(x \mid \theta^*_j) \Big)^2, \end{aligned} \tag{42}$$

where θ_j^* is the best θ_j from M_j (but not necessarily best from M_{j^*}). So, as $n \to \infty$, we get that

$$E_Y(Y(x) - \hat{Y}_{BMA}(x))^2 \leq E_Y(Y(x) - f_j(x \mid \hat{\theta}_j))^2. \tag{43}$$

from (41) and (42), pointwise in x, for each fixed j, giving the theorem. ■

For expression (43) to hold requires that $f_T(x) = f_{j^*}(x \mid \theta_{j^*}^*)$ for some specific optimal $\theta_{j^*}^*$. If that fails then the best (wrong) model is the one indexed by the θ_{j^*} value that gives a distribution closest to P_T in relative entropy, not necessarily the same as in squared error, as first identified in Berk (1966). If we assume that the posterior weights are well-behaved, we obtain the following extension of Theorem 5 to the wrong model setting.

Corollary 1 *Let $Y = f_T(x) + \epsilon$ and let $j^* = \arg\min_j D(f_T \| f_j(\theta_j))$ and assume the other hypotheses of Prop.1. If $W(M_{j^*} \mid \mathcal{D}) \to 1$, then*

$$\int E(Y(x) - \hat{Y}_{BMA}(x))^2 \mathrm{d}x \leq \int E(Y - f_j(x|\hat{\theta}_j))^2 \mathrm{d}x.$$

Proof Rewrite the proof of Theorem 5 using f_T in place of f_{j^*}. ■

Some comments on more general and verifiable hypotheses for Cor. 1 are worth making. These elaborate on the Remark after Prop. 1 in view of the proofs of that result and Theorem 5. If the setting is that the true model is M_T not one of the M_j's then convergences of quantities derived from any M_j must be assessed in M_T, about which we know little. By contrast, in the results so far, we assessed convergence in M_{j^*}. To generalize Prop. 1, we require all of its assumtions plus Lemma 3. In principle, Lemma 3 permits us to generalize the main result in Clarke and Barron (1988) to the wrong model INID case so that the wrong model INID version of (34) will hold. Likewise, Lemma 3 can be used to generalize Lemma 1 to get a generalization of Wilks theorem to control (36). Since the leading term $\mathcal{O}(n)$ is unchaged we get a more general version of Prop. 1 that gives the convergence of the posterior weights as needed in Cor. 1. We also get the more general statement of Prop. 1 for use in a wrong-model INID version of Theorem 5 mentioned in the proof of Cor. 1 where we noted the proof is unchanged apart from using M_T and a different j^*.

To conclude this section, we note that if p is allowed to increase slowly in the $f_j(x \mid \theta_j)$'s then as $n \to \infty$ all of the results continue to hold.

4. The Stacking Model Average Predictor

Stacking was first introduced by Wolpert (1992) and studied primarily as a predictor in numerous contexts such as regression Breiman (1996a), Clarke (2003), Sill et al. (2009), classification Ting and Witten (1999), Ozay and Vural (2012), and density estimation Smyth and Wolpert (1999). Stacking has also been used to estimate error rates Rokach (2010).

Stacking has usually been seen as a frequentist procedure. However, Clyde and Iversen (2013) explicitly extended stacking to $\mathcal{M}$-open problems, brought it into the Bayes paradigm, and examined the effect of varying the constraints on α in the optimization (44). Le and

Clarke (2017) formally showed that stacking can be regarded as the Bayes action under several loss functions, asymptotically.

The basic idea is that the models in $\mathcal{F}$ can be usefully combined to give the predictor

$$\hat{Y}_{stack}(x) = \sum_{j=1}^{J} \hat{\alpha}_j \hat{f}_j(x).$$

The $\hat{\alpha}_j$'s are obtained by invoking an optimality property similar to cross-validation. More formally, let $\hat{f}_{j,-i}$ be the estimate of f_j using $n-1$ of the n data points and dropping the i-th one. Then the estimated weight vector $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_J)$ is

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \alpha_j \hat{f}_{j,-i}(x_i) \right)^2 . \tag{44}$$

Expression (44) corresponds to leave-one-out cross-validation but can be modified to correspond to leave-K-out cross-validation. Also, $\alpha \in \mathbb{R}^J$ but may be restricted, e.g. to $\mathbb{R}^{+J}$ or the simplex in $\mathbb{R}^J$.

In this section we establish the same three kinds of results for stacking as we described in general for model averages in Sec. 1 and established for MMA in Sec. 2. Recall, these are consistency of parameter estimates, an oracle style inequality for the empirical risk, and better asymptotic performance of the model average predictor than its component predictors. We begin by showing these results for fix p and then state the extension of the results for increasing p.

For each $j = 1, \ldots, J$ let $p_j(\cdot \mid \theta_j)$ be the density corresponding to M_j, equipped with a prior density $w_j(\theta_j)$ leading to the posterior density $w_j(\theta_j \mid \mathcal{D}_n)$. We assume $\theta_j \in \Theta_j \subset \mathbb{R}^{d_j}$ and that each Θ_j is open with compact closure satisfying $\bar{\Theta}_j^o = \Theta_j$. First, we have the following theorem. Here we assume $x \in \mathcal{X}$, a compact set, and that it is the x-values that make the $p_j(y \mid x, \theta_j)$ independent but not identical (INID).

Our consistency result is the following.

Theorem 6 *: Assume $Y_i = f_T(x_i \mid \theta) + \epsilon_i$, where the ϵ_i's are IID mean zero and variance σ^2 and for all x $f_T(x \mid \theta) = \sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta_e)$ with $\theta = (\theta_1, \ldots, \theta_J)$, where for each fixed set of θ_j's, θ_e is a function of $\theta_1, \ldots, \theta_J$, the f_j's are orthonormal (under F), continuous, bounded on their domain, orthogonal to e_J in L^2, and $e_J(x \mid \theta_e) \to 0$ as $J \to \infty$. Write the stacking predictor for step $n+1$ as $\hat{Y}_{stack}(x_{n+1}) = \sum_{j=1}^{J} \hat{\alpha}_j f_j(x_{n+1} \mid \hat{\theta}_j)$ where the α_f's are as in (44). Then, provided each $\hat{\theta}_j$ is consistent for θ_j in p_T, the distribution of Y, as $n \to \infty$ we have that, for all $j = 1, \ldots, p$*

$$\hat{\alpha}_j \to \alpha_j \quad in \quad p_T. \tag{45}$$

Remark 1: The basic technique of proof extends to the case of constrained α_j's but becomes more complicated because it uses Theorem 3.1 (or Corollary 3.1) in Le and Clarke (2017) rather than Theorem 3.2.

Remark 2: Sufficient conditions for the consistency of the $\hat{\theta}_j$'s are given in Lemma 1 and Lemma 2 in Appendix A.1 and Lemma 3 and Lemma 4 in Appendix A.2.

Proof : Theorem 3.2 in Le and Clarke (2017) gives that the stacking weights $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_J)$ achieving

$$\min_{\alpha} \sum_{i=1}^{n} \left(y_i(x_i) - \sum_{j=1}^{J} \alpha_j \hat{y}_{j,-i}(x_i) \right)^2$$

are of the form

$$\hat{\alpha} = T_n^{-1} c, \tag{46}$$

where

$$\begin{aligned} T_n &= \frac{1}{n} \left(\sum_{i=1}^{n} \hat{y}_{l,-i}(x_i) \hat{y}_{j,-i}(x_i) \right)_{J \times J}, \\ c &= \frac{1}{n} \left(\sum_{i=1}^{n} y_i(x_i) \hat{y}_{1,-i}(x_i), \cdots, \sum_{i=1}^{n} y_i(x_i) \hat{y}_{J,-i}(x_i) \right)' \end{aligned} \tag{47}$$

and $\hat{y}_{l,-i}(x_i) = f_l(x_i \mid \hat{\theta}_l(\hat{x}_i)$ and $\hat{x}_i$ means that x_i is omitted, for $\ell = 1, \dots, J$.

To see that the $\hat{\alpha}_j$'s are consistent for the true values α_j, begin by writing

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \hat{y}_{l,-i}(x_i) \hat{y}_{j,-i}(x_i) &= \frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \hat{\theta}_l(\hat{x}_i)) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} \Big(f_l(x_i \mid \hat{\theta}_l(\hat{x}_i)) - f_l(x_i \mid \theta_l) \Big) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \theta_l) \Big(f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) - f_j(x_i \mid \theta_j) \Big) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \theta_l) f_j(x_i \mid \theta_j). \end{aligned} \tag{48}$$

By the consistency of $\hat{\theta}(\hat{x}_i)$ for any i, the dominated convergence theorem gives that the first and second terms in the right hand side of (48) go to 0. Thus,

$$\frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \hat{\theta}_l(\hat{x}_i)) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) = \frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \theta_l) f_j(x_i \mid \theta_j) + o_P(1), \tag{49}$$

and by the law of large numbers

$$\frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \theta_l) f_j(x_i \mid \theta_j) \to \int f_l(x \mid \theta_l) f_j(x \mid \theta_j) dF(x). \tag{50}$$

So, for $\hat{Y}_{stack}$ formed from an orthonormal basis as before, we have that

$$\frac{1}{n} \sum_{i=1}^{n} f_l(x_i \mid \hat{\theta}_l(\hat{x}_i)) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) = \begin{cases} o_P(1), & \text{if } l \neq j \\ 1 + o_P(1), & \text{if } l = j \end{cases}, \tag{51}$$

and therefore

$$T_n = (1 + o_P(1))Id_J. \tag{52}$$

Using (52) with (47) we get

$$\hat{\alpha} = (1 + o_P(1))Id_J * c, \tag{53}$$

since $1/(1 + o_P(1)) = 1 + o_P(1)$.

From (47), the typical entry in c is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} y_i(x_i)\hat{y}_{j,-i}(x_i) &= \frac{1}{n}\sum_{i=1}^{n}(f_T(x_i \mid \theta_T) + \epsilon_i) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_T(x_i \mid \theta_T) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)).\end{aligned} \tag{54}$$

We see that the second term in (54) is $o_P(1)$ by writing it as

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(x_i \mid \theta_j) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i \left[f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) - f_j(x_i \mid \theta_j)\right].$$

The first sum has IID terms since the $X_i \sim F$ are IID and independent of the ϵ_i's that are also IID. The limit of the sum is zero a.s. since the expectation of each ϵ_i is zero. The second sum is also $o_P(1)$ but there are a few more steps to the argument. Write $U_i^n = f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) - f_j(x_i \mid \theta_j)$ and note U_i^n is independent of ϵ_i. By Markov's inequality, for any $\epsilon > 0$ we have

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i U_i^n\right| \geq \epsilon\right) &\leq \frac{E\left|\epsilon_1\right|}{n\epsilon}\sum_{i=1}^{n} E\left|U_i^n\right| \\ &\leq \frac{E\left|\epsilon_1\right|}{n\epsilon}\sum_{i=1}^{n} E\left|U_i^n\right|\chi_{\{|U_i^n|\leq\eta\}} + \frac{E\left|\epsilon_1\right|}{n\epsilon}\sum_{i=1}^{n} E\left|U_i^n\right|\chi_{\{|U_i^n|>\eta\}} \\ &\leq \frac{E\left|\epsilon_1\right|}{\epsilon}\eta + \frac{E\left|\epsilon_1\right|}{n\epsilon}\sum_{i=1}^{n}\sqrt{E\left|U_i^n\right|^2 P(|U_i^n| > \eta)} \end{aligned} \tag{55}$$

for any $\eta > 0$, using Cauchy's inequality in the last step. The summands in (55) are a product of terms that are bounded or $o_P(1)$. Accordingly, (55) is $o_P(1)$ and so is the second sum. So, (54) is

$$\frac{1}{n}\sum_{i=1}^{n} f_T(x_i \mid \theta_T) f_j(x_i \mid \hat{\theta}_j(\hat{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} f_T(x_i \mid \theta_T) f_j(x_i \mid \theta_j) + o_P(1). \tag{56}$$

Now as $n \to \infty$ we get

$$\frac{1}{n}\sum_{i=1}^{n} f_T(x_i \mid \theta_T) f_j(x_i \mid \theta_j) = \int f_T(x \mid \theta_T) f_j(x \mid \theta_j) dF(x). \tag{57}$$

Using (57) in (53) gives that

$$\hat{\alpha}_j \overset{P}{\to} \int f_T(x \mid \theta_T) f_j(x \mid \theta_j) dF(x) = \alpha_j. \tag{58}$$

So, in the absence of constraints the estimated stacking coefficients are consistent for the Fourier coefficients of f_T with respect to the orthonormal functions being stacked. ■

Next we turn to the empirical risk. Since the stacking coefficients are derived from a condition similar to cross-validation we will see that dropping one data point does not affect convergence of predictors. Indeed, the proof can be extended to show that dropping any finite number of data points does not affect convergence thereby permitting leave-k-out cross-validation if desired.

Our first result states and proves a result on dropping data points from Bayes predictors posterior means. Only the latter clause is used in the sequel however we think the fist clause is of independent interest.

Proposition 2 *: Assume the conditions of Lemma 1 and Lemma 2.*
Clause I: If, for each $j = 1, \ldots, J$, the Bayes predictor

$$\hat{f}_j(x_{n+1}) = E_j(Y_{n+1} \mid \mathcal{D}_n) = \int\int y_{n+1} p_j(y_{n+1} \mid x_{n+1}, \theta_j) w_j(\theta_j \mid \mathcal{D}_n) d\theta_j dy_{n+1}$$

is used to generate predictions at the $n+1$ step and the $E_j Y_{n+1}$'s are uniformly bounded as functions of x over all values of j and n, then

$$\sup_{x \in \mathcal{X}} \left[\hat{f}_j(x) - \hat{f}_{j,-i}(x)\right] = \sup_{x \in \mathcal{X}} \left[E_j(Y_{n+1} \mid \mathcal{D}_n) - E_j(Y_{n+1} \mid \mathcal{D}_{n,-i})\right] \to 0$$

almost everywhere in model M_j as $n \to \infty$, where $\mathcal{D}_{n,-i} = \mathcal{D}_n \setminus \{(x_i, y_i)\}$.
Clause II: The same results hold if the posterior expectation of Y_{n+1} is replaced by Θ_j, i.e., $E_j(Y_{n+1} \mid \mathcal{D}_n)$ is replaced by $E_j(\Theta_j \mid \mathcal{D}_n)$ and similarly for $\mathcal{D}_{n,-i}$.

Proof : Let (x_{n+1}, Y_{n+1}) be a new data point. For any fixed model index j, write

$$\hat{f}_{j,-i}(x_{n+1}) = E_j(Y_{n+1} \mid \mathcal{D}_{n,-i}) = \int\int y_{n+1} p_j(y_{n+1} \mid x_{n+1}, \theta_j) w_j(\theta_j \mid \mathcal{D}_{n,-i}) d\theta_j dy_{n+1}.$$

Also note that

$$\begin{aligned} w_j(\theta_j \mid \mathcal{D}_n) &= w_j(\theta_j \mid (x_1, y_1), \ldots, (x_n, y_n)), \\ w_j(\theta_j \mid \mathcal{D}_{n,-i}) &= w_j(\theta_j \mid (x_1, y_1), \ldots, \widehat{(x_i, y_i)}, \ldots, (x_n, y_n)), \end{aligned}$$

where the caret $\widehat{\ }$ means the term is deleted. Therefore,

$$\begin{aligned} &|\hat{f}_j(x_{n+1}) - \hat{f}_{j,-i}(x_{n+1})| \\ &\leq \int |y_{n+1}| \int \left[p_j(y_{n+1} \mid x_{n+1}, \theta_j) \left|w_j(\theta_j \mid \mathcal{D}_n) - w_j(\theta_j \mid \mathcal{D}_{n,-i})\right|\right] d\theta_j dy_{n+1} \end{aligned} \tag{59}$$

Let $\hat{\theta}_{n,j}$ be the MLE of θ_j defined in Lemma 1. For $x_{n+1} \in \mathcal{X}$ (compact) and $y_{n+1} \in \mathcal{Y}$ (compact), since $p_j(y_{n+1} \mid x_{n+1}, \theta_j)$ is bounded, Lemma 2 gives that the right hand side of (59) is bounded by

$$\begin{aligned}
&M \int \left| w_j(\theta_j \mid \mathcal{D}_n) - N\left(\hat{\theta}_{n,j}, \left[\sum_{i=1}^{n} I_i(\theta_{0j} \mid x_i)\right]^{-1}\right)\right| d\theta_j \\
&+ M \int \left| N\left(\hat{\theta}_{n,j}, \left[\sum_{i=1}^{n} I_i(\theta_{0j} \mid x_i)\right]^{-1}\right) - N\left(\hat{\theta}_{n,j,-i}, \left[\sum_{k \neq i} I_k(\theta_{0j} \mid x_k)\right]^{-1}\right)\right| d\theta_j \\
&+ M \int \left| N\left(\hat{\theta}_{n,j,-i}, \left[\sum_{k \neq i} I_k(\theta_{0j} \mid x_k)\right]^{-1}\right) - w_j(\theta_j \mid \mathcal{D}_{n,-i})\right| d\theta_j,
\end{aligned}$$

which does not depend on x_{n+1} for some constant M. Hence,

$$\begin{aligned}
&\sup_{x_{n+1} \in \mathcal{X}} |\hat{f}_j(x_{n+1}) - \hat{f}_{j,-i}(x_{n+1})| \\
&\leq M \int \left| w_j(\theta_j \mid \mathcal{D}_n) - N\left(\hat{\theta}_{n,j}, \left[\sum_{i=1}^{n} I_i(\theta_{0j} \mid x_i)\right]^{-1}\right)\right| d\theta_j \\
&+ M \int \left| N\left(\hat{\theta}_{n,j}, \left[\sum_{i=1}^{n} I_i(\theta_{0j} \mid x_i)\right]^{-1}\right) - N\left(\hat{\theta}_{n,j,-i}, \left[\sum_{k \neq i} I_k(\theta_{0j} \mid x_k)\right]^{-1}\right)\right| d\theta_j \qquad (60) \\
&+ M \int \left| N\left(\hat{\theta}_{n,j,-i}, \left[\sum_{k \neq i} I_k(\theta_{0j} \mid x_k)\right]^{-1}\right) - w_j(\theta_j \mid \mathcal{D}_{n,-i})\right| d\theta_j,
\end{aligned}$$

Using Scheffé's theorem and Lemma 2 gives convergence in L^1 for the first and third terms in (60), since by Lemma 1 $\hat{\theta}_{n,j} - \hat{\theta}_{n,j,-i} \overset{a.s.}{\to} 0$. For the middle term of (60), note that $\sum_{i=1}^{n} I_i(\theta_{0j} \mid x_i) \to \infty$ and $\sum_{k \neq i} I_k(\theta_{0j} \mid x_k) \to \infty$. Now, Lemma 1 again implies then the middle term goes to 0. ■

Remark: This result can be generalized to the case that $p_T \notin \mathcal{F}$ by using $\hat{\theta}_{n,j}$ in a more complicated argument starting at (59) and using Lemma 4. For actual usage, sufficient conditions for the $E_j Y_{n+1}$'s be uniformly bounded in j and n can be obtained by assuming x varies over a compact set, by assuming that the tail behavior of the f_j's in terms of x is properly controlled, by assuming that the range of θ_j is restricted, or by a combination of such assumptions.

Observe that the cumulative predictive loss of the stacking average under squared error loss is

$$L_n(Stacking) = \sum_{i=1}^{n} (y_i - \hat{Y}_{stack}(x_i))^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} \hat{\alpha}_j \hat{f}_j(x_i))^2,$$

and the cumulative predictive loss from using individual model M_j is

$$L_n(M_j) = \sum_{i=1}^{n} (y_i - \hat{f}_j(x_i))^2.$$

So, we can use Prop. 2 and Lemma 5 to get a result for the $\mathcal{M}$-complete and -closed cases. For any j, the definition of the stacking coefficients gives

$$\begin{aligned} & L_n(Stacking) \\ & = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} \hat{\alpha}_j \hat{f}_{j,-i}(x_i))^2 \leq \sum_{i=1}^{n} (y_i - \hat{f}_{j,-i}(x_i))^2. \end{aligned} \tag{61}$$

When a problem is $\mathcal{M}$-complete or -closed, the true model exists even if it's inaccessible, and hence can be used to define a mode of convergence. Using the posterior means as estimators for the θ_j's, write

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{f}_{j,-i}(x_i))^2 = & \left[\sum_{i=1}^{n} (y_i - \hat{f}_{j,-i}(x_i))^2 - \sum_{i=1}^{n} (y_i - \hat{f}_j(x_i))^2 \right] \\ & + \sum_{i=1}^{n} (y_i - \hat{f}_j(x_i))^2, \end{aligned} \tag{62}$$

and consider the differences

$$\begin{aligned} & (y_i - \hat{f}_{j,-i}(x_i))^2 - (y_i - \hat{f}_j(x_i))^2 \\ & = (\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i)) \\ & = f_j'(x_i \mid \hat{\theta}_j^*) \left(E_j(\Theta \mid \mathcal{D}_n) - E_j(\Theta \mid \mathcal{D}_{n,-i}) \right) \left(2y_i - f_j(x_i \mid E_j(\Theta \mid \mathcal{D}_n)) - f_j(x_i \mid E_j(\Theta \mid \mathcal{D}_{n,-i}) \right), \end{aligned} \tag{63}$$

where we have used a first order Taylor expansion in the first factor. It is easy to see that the first factor in (63) is bounded since the f_j's are continuous functions on compact sets. The second factor in (63) goes to zero in probability by the same proof as Prop. 2 applied to the posterior means of θ_j's rather than Y_i's. Now, using Lemma 5 for parameters rather than Y's, on (63) and then applying Lemma 6 gives us the following bound on the pointwise error of stacking.

Theorem 7 *: Assume the hypotheses of Proposition 2 and Lemmas 5 for parameters and Lemma 6. For $j = 1, \ldots, J$, when M_j is true, as $n \to \infty$, we have*

$$L_n(Stacking) \leq L_n(M_j) + o_P(n), \tag{64}$$

Remark: In fact, if we use $\sqrt{n}(E_j(\Theta \mid \mathcal{D}_n) - E_j(\Theta \mid \mathcal{D}_{n,-i})) = \mathcal{O}_P(1)$, as is easy to show by standard techniques, see the proof of Prop. 2 or Clarke (1989) p. 71, Theorem 7 can be readily improved to

$$L_n(Stacking) \leq L_n(M_j) + O_P\left(1/\sqrt{n}\right).$$

Proof : Lemma 5 for parameters, see (63), shows that $(y_i - \hat{f}_{j,-i}(x_i))^2 - (y_i - \hat{f}_j(x_i))^2 = (\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i)) = o_P(1)$ uniformly. Then, by Lemma 6 and Proposition 2 for parameters the first term on the right of (62) is $o_P(n)$ in model M_j as $n \to \infty$ and the second term is $L_n(M_j)$ for any $j = 1, \ldots, J$. Therefore,

$$L_n(Stacking) \leq L_n(M_j) + o_P(n),$$

as $n \to \infty$ for $j = 1, \ldots, J$ when M_j is true. ■

Remark: The introduction of new parameters α_j for $j = 1, \ldots, J$ can in principle lead to a stacking average that overfits the data in which case generalisation would be poor. However, Breiman (1996a) advocates choosing the f_j's to be as different from each other as possible and here we are assuming n large. In these settings, any overfitting will typically be slight.

Expression (64) is evidence that stacking provides no worse prediction on average, under squared error loss, than using the predictor from any single model in the stacking average. However, our result is in cumulative error over n predictions, not the error of an individual prediction. It is usually easier to prove results such as (64) rather than their prequential versions even though cumulative prequential loss is usually smaller than regular cumulative loss. For prequential results in $\mathcal{M}$-complete and $\mathcal{M}$-closed cases see Dawid (1984), Skouras and Dawid (1998), Dawid and Vovk (1999), and Skouras and Dawid (1999). In many such cases, ordering the performance of methods by prequential cumulative loss is the same as ordering of methods by regular cumulative loss. So, sometimes it may be satisfactory to compare predictors in $\mathcal{M}$-open cases using regular cumulative loss.

To address individual predictions, we state and prove a result in the same spirit as (64) but for $\mathcal{M}$-complete and $\mathcal{M}$-closed DG's. When we take expectations over X we denote the distribution function by F.

Theorem 8 *: Assume $Y_i = f_T(x_i \mid \theta) + \epsilon_i$, where the ϵ_i's are IID mean zero and variance σ^2 and for all x $f_T(x \mid \theta) = \sum_{j=1}^J \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta_e)$ with $\theta = (\theta_1, \ldots, \theta_J)$, where for each fixed set of θ_j's, θ_e is a function of $\theta_1, \ldots, \theta_J$, the f_j's are orthonormal (under F), continuous, bounded on their domain, orthogonal to e_J in L^2, and $e_J(x \mid \theta_e) \to 0$ as $J \to \infty$. Write the stacking predictor for step $n+1$ as $\hat{Y}_{stack}(x_{n+1}) = \sum_{j=1}^J \hat{\alpha}_j f_j(x_{n+1} \mid \hat{\theta}_j)$ where the α_j's are as in (44). Then, provided each $\hat{\theta}_j$ is consistent for θ_j in p_T, the distribution of Y, as $n \to \infty$ we have*

$$\limsup_{n\to\infty} E_Y(Y(x) - f_j(x \mid \hat{\theta}_j))^2 - E_Y(Y(x) - \hat{Y}_{stack}(x))^2 \geq 0 \tag{65}$$

pointwise in x, and, by the properties of e_J and the f_j's,

$$\limsup_{n\to\infty} \left(\int E_Y(Y(x) - f_j(x \mid \hat{\theta}_j))^2 dx - \int E_Y(Y(x) - \hat{Y}_{stack}(x))^2 dx \right) \geq 0. \tag{66}$$

Remarks: (i) One common choice for the $\hat{\theta}_j$'s are the posterior means of the θ_j's. (ii) Theorem 8 means that the stacking predictor is closer to Y than the plug-in predictor from any individual model M_j is.

Proof : It is enough to compare the squared error of stacking with model j. Our hypotheses give that

$$f_T(x) = \sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta), \tag{67}$$

where $e_J \perp f_j$ for all j and $e_J(x \mid \theta) \to 0$ in L^2 as $J \to \infty$, and

$$\hat{Y}_{stack}(x) = f_{stack}(x) = \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j). \tag{68}$$

Also, we can assume that the $\hat{\alpha}_j$'s are in a compact set. Now,

$$\begin{aligned} E_Y(Y(x) - f_{stack}(x))^2 &= E_Y\left(Y(x) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)^2 \\ &= E_Y\left(Y(x) - EY(x) + EY(x) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)^2 \\ &= E_Y\left(Y(x) - EY(x)\right)^2 + E_Y\left(EY(x) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)^2 \\ &\quad + 2E_Y\left[(Y(x) - EY(x))\left(EY(x) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)\right]. \end{aligned}$$

Since the difference $Y(x) - EY(x) = \epsilon_{n+1}$ with $E(\epsilon_{n+1}) = 0$ and independent of earlier data i.e. for $i \leq n$, we have

$$\begin{aligned} E_Y(Y(x) - f_{stack}(x))^2 &= \sigma^2 + E_Y\left(\sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)^2 \\ &= \sigma^2 + e_J^2(x \mid \theta) + E_Y\left(\sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) - \sum_{j=1}^{J} \hat{\alpha}_j f_j(x \mid \hat{\theta}_j)\right)^2. \end{aligned}$$

Using consistency of $\hat{\theta}_j$'s and $\hat{\alpha}_j$'s, the dominated convergence theorem gives

$$\begin{aligned} E_Y(Y(x) - f_{stack}(x))^2 &= \sigma^2 + e_J^2(x \mid \theta) + o_P(1) \\ &= \sigma^2 + o_P(1), \end{aligned} \tag{69}$$

for $J = J_n \to \infty$, i.e., for n large enough.

Next, for any individual model M_k, similar arguments give

$$\begin{aligned} E_Y(Y(x) - f_k(x \mid \hat{\theta}_k))^2 &= \sigma^2 + E_Y \left(\sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta) - f_k(x \mid \hat{\theta}_k) \right)^2 \\ &= \sigma^2 + \left(\sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta) - f_k(x \mid \theta_k) \right)^2 \\ &= \sigma^2 + \left(\sum_{j=1}^{J} (\alpha_j - 1_{k,j}) f_j(x \mid \theta_j) + e_J(x \mid \theta) \right)^2 \\ &= \sigma^2 + \left(\sum_{j=1}^{J} (\alpha_j - 1_{k,j}) f_j(x \mid \theta_j) \right)^2 + o_P(1) \qquad (70) \end{aligned}$$

for $J = J_n \to \infty$, i.e., for n large enough, where $1_{k,j} = 1$ if $k = j$ and 0 otherwise.

Therefore, in the limit, the right hand side of (70) bigger than the right hand side of (69) for any x with equality if and only if there exists an j such that $f_j = f_T$. So, (65) follows. Now, expression (66) follows by the dominated convergence theorem. ■

To conclude this section we consider increasing dimension of x and θ. Let us reinterpret the notation from Theorem 8 to mean that in $Y_i = f_T(x_i \mid \theta) + \epsilon_i$ and $f_T(x \mid \theta) = \sum_{j=1}^{J} \alpha_j f_j(x \mid \theta_j) + e_J(x \mid \theta_e)$ we mean that f_j for $j \leq J = J_p$ has arguments x and θ_j with $\dim(x), \dim(\theta_j) \leq p$ and any dependence on dimensions $p+1, p+2, \ldots$ is in $e_J(x \mid \theta_e)$. Then it is easy to verify that for $p \to \infty$ slow enough the corresponding statements of the results in this section remain true. In particular, Theorems 6, 7 and 8 continue to hold in the limit as $p = p_n \to \infty$ slowly with n as $n \to \infty$.

5. Bagging

A fourth model averaging technique is bagging ('bootstrap aggregating'). It was introduced by Breiman (1996b) as a general strategy to improve the precision of model-based predictors. Usually, the model is thought to be good in the sense of being unbiased but gives predictors that are highly variable so that the bagging procedure will stabilize them.

Originally, bagging was proposed in the context of trees for $\mathcal{M}$-complete DG's in classification problems. In this context, Breiman (2001) established several optimality properties of bagging. Bagging has also been used, usually without comment, for regression problems that are $\mathcal{M}$-open as well. For instance, Strobl et al. (2009) provide several examples that seem $\mathcal{M}$-open rather than $\mathcal{M}$-complete as well as a discussion of the key features of bagging in practice. However, there remains relatively little understanding how bagging works apart from the results in Breiman (1996a) and Buhlmann and Yu (2002). Here, our main addition to the conceptual understanding of bagging is to show it is, asymptotically, a special case of a pseudo-BMA.

The bagging model average is defined as follows. Given a sample, fit a model $\hat{f}(\cdot)$ and consider predicting the response for a new value of x. Usually, $\hat{f}(x) = f(x \mid \hat{\theta})$ for some estimator $\hat{\theta}$. A bagged predictor for Y at x is found by drawing B bootstrap samples of size n, say $\mathcal{D}_b$ from $\mathcal{D}_n$ for $b = 1, \ldots, B$, using each $\mathcal{D}_b$ to produce an $\hat{f}_b(x)$, and taking the prediction to be the average

$$\hat{Y}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x). \tag{71}$$

As in earlier sections, we begin with a consistency result for bagging. This is easy because the consistency of $\hat{f}$ gives the consistency for $\hat{Y}_{bag}$ as in the following.

Proposition 3 *: Suppose $Y = f(x \mid \theta) + \epsilon$ with ϵ mean zero random error, θ a finite dimensional real parameter, f continuously differentiable in its arguments, and that we have independent data $\mathcal{D}_n$. If there is an estimator $\hat{\theta}$ that is $\sqrt{n}$-consistent for θ then $\hat{f}_{bag}(\cdot)$ in* (71) *is consistent for $f(\cdot \mid \theta)$ pointwise in x as well.*

Remark: The existence of a $\sqrt{n}$-consistent estimator is guaranteed by the hypotheses of Lemma 1 for the MLE or Lemma 2 for Bayes estimators. To save on technicality, we only give an outline of the proof. To get the result formally would require an argument on subsequences of $f(x \mid \hat{\theta}_{b,n})$ as b and n slowly increase. This has been omitted for brevity.
Proof (outline): We set up an application of the strong law of large numbers established for dependent variables in Theorem 1 in Hu et al. (2008). There are three hypotheses.

First, we must bound the covariances of the terms in (71). Let $\hat{\theta}_{b,n}$ and $\hat{\theta}_{b,n'}$ be values of the $\sqrt{n}$ consistent estimator for bootstrap sample b from $\mathcal{D}_n$ and $\mathcal{D}_{n'}$ where $n' = n + k$. Since the data sets $\mathcal{D}_b$ are chosen randomly from $\mathcal{D}_n$ and $\mathcal{D}_{n'}$, and n is increasing, every $\hat{f}(x \mid \hat{\theta}_b)$ is $\sqrt{n}$-consistent for $f(x \mid \theta)$, pointwise in x. So, with high probability $\hat{\theta}_{b,n}$ and $\hat{\theta}_{b,n'}$ are close to θ. So, we can use Taylor expansions at θ:

$$\begin{aligned} f(x \mid \hat{\theta}_{b,n}) &= f(x \mid \theta) + f'(x \mid \theta^*_n)(\hat{\theta}_{b,n} - \theta) \\ f(x \mid \hat{\theta}_{b,n'}) &= f(x \mid \theta) + f'(x \mid \theta^*_{n'})(\hat{\theta}_{b,n'} - \theta), \end{aligned}$$

where θ^*_n and $\theta^*_{n'}$ are on the lines joining their respective bootstrap estimate and θ.

The first hypothesis is satisfied: Let $C > 0$ denote a positive constant not necessarily the same from occurrence to occurrence. Write:

$$\begin{aligned} |\mathrm{COV}(f(x \mid \hat{\theta}_{b,n}), f(x \mid \hat{\theta}_{b,n'}))| &\leq C(f)\mathrm{COV}((\hat{\theta}_{b,n} - \theta)(\hat{\theta}_{b,n} - \theta)) \\ &\approx \frac{C(f)}{.62\sqrt{n(n+k)}} E(\sqrt{.62n}(\hat{\theta}_{b,n} - \theta)\sqrt{.62n}(\hat{\theta}_{b,n} - \theta)) \\ &\overset{\infty}{\cong} \frac{C(f)}{.62\sqrt{n(n+k)}} \sigma_{n,n'} = \frac{C(f,\sigma)}{\sqrt{n(n+k)}}, \end{aligned}$$

where .62 is the approximate fraction a bootstrap sample represents from the full sample and $\sigma_{n,n'}$ is the covariance indicated. The right hand side shows that

$$\rho_k = \sup_{n \geq 1} |\mathrm{COV}(f(x \mid \hat{\theta}_{b,n}), f(x \mid \hat{\theta}_{b,n'}))| = \mathcal{O}(1/\sqrt{k}),$$

which is the first hypothesis of Theorem 1.

The second hypothesis is on the asymptotic rate of $f(x \mid \hat{\theta}_{b,n})$. We see that

$$\frac{\mathrm{VAR}(f(x \mid \hat{\theta}_{b,n})}{n^2} = \mathcal{O}(1/n^3)$$

which is smaller than the largest bound on the variance that Theorem 1 requires.

The third hypothesis is that ρ_k/k^γ be summable for some $\gamma \in (0,1)$. Clearly, choosing $\gamma = .75$ suffices. So, the hypotheses of Theorem 1 in Hu et al. (2008) are satisfied.

Now Theorem 1 gives the Proposition: As $n \to \infty$ and $B = B_n \to \infty$,

$$\sum_{b=1}^{B} \frac{f(x \mid \hat{\theta}_{b,n})}{n} \overset{a.s.}{\to} Ef(x \mid \hat{\theta})) \overset{a.s.}{\to} f(x \mid \theta).$$

■

The second sort of result for a model average procedure is an oracle type inequality or bound on the empirical risk as discussed in Sec. 1. For aggregation procedures such as bagging there is only one model and it is the data that is re-used. Cross-validation procedures have a similar property but use disjoint subsets of the data making results more feasible. At this time of writing, there seems to be no oracle inequalities or bounds on empirical risks for bagging in general. Partial results are available in special cases; see Dalalyan and Salmon (2012), Montuelle and Pennec (2018) and Maillard et al. (2021). However, even these are very difficult to establish—perhaps because the terms in the bagging predictor are symmetric in the data and hence have equal standing probabilistically. So, here we start by observing that Breiman's Theorem (recorded here as Proposition 5 in Sec. 6) provides a bound for bagging in the same spirit as an oracle inequality would even though it is phrased in terms of risk rather than empirical risk.

The third sort of result for model averaging is to verify that the model average performs no worse asymptotically than any of its components. Because all the terms in a bagging predictor are probabilistically equivalent, we show this in a different sense. We first show that a bagging predictor is a sort of pseudo-BMA and the second identifies the term in the pseudo-BMA that gives the best prediction. This is somewhat like an oracle inequality but is no substitute for results treating the empirical risk itself. On the other hand, this will mean that some of the benefits of BMA carry over to bagging asymptotically. Thus, our strategy is to prove a variant of Prop. 1 for bagging, starting with unidimensional θ, extending to multidimensional θ, and then decomposing a bagging predictor into terms from the pseudo-BMA.

We begin by defining $f(x \mid \theta)$ be a bounded continuous function of (x, θ), $x \in \mathbb{R}^p$, $\theta \in \Theta$ an open interval in $\mathbb{R}$; we will generalize to d-dimensional θ shortly. Assume the data $\mathcal{D} = \{(x_i, y_i)\}|_{i=1}^n$ are independent over i and come from

$$Y(x) = f(x \mid \theta_T) + \epsilon. \tag{72}$$

Note that in this section, we do not consider a wrong model analysis because, once p is large enough—as is the case with, say, bagging trees, the bagged model is essentially nonparametric.

To state our first result, consider an approximation to the BMA formed by partitioning the parameter space into K disjoint and exhaustive (measureable) sets $A_1, \ldots, A_K$, taken to be intervals in the unidimensional parameter case. Write $I_b = A_k \Leftrightarrow \hat{\theta}_b \in A_k$, where the $\hat{\theta}_b$'s are the estimates of θ from the B bootstrap samples.

Suppose the MLE $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is consistent for θ_T under P_T as defined in (72). For any x, the bagging predictor based on f is now

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f(x \mid \hat{\theta}_b) \tag{73}$$

where $\hat{f}_b(x) = f(x \mid \hat{\theta}_b)$ and $\hat{\theta}_b$ is the value of the estimator on the b-th of B bootstrap samples from $\mathcal{D}$. Fix C and let $\hat{F} = \hat{F}_{n,B}(\cdot)$ be the empirical distribution function generated by the $\hat{\theta}_b$'s. For $c = 0, \ldots, C$, let $\{\alpha_c\}$ be the ℓ/C quantiles of $\hat{F}$ for $\ell = 0, 1, \ldots, C$ (α_0 can be taken as $-\infty$ if x is not bounded below). Let $w(\cdot)$ be the density of θ on Θ assumed to satisfy $w(\theta) > \gamma > 0$ on Θ. When needed, we write the probability measure generated from w as $W(\cdot)$. Write an empirical form of the BMA as

$$\hat{Y}_{BMA}(x) = \sum_{c=1}^{C} W([\alpha_{c-1}, \alpha_c] \mid \mathcal{D}) E(Y \mid M_c, \mathcal{D}),$$

where M_c is the (empirical) submodel $\{f(\cdot \mid \theta) \mid \theta \in [\alpha_{c-1}, \alpha_c]\}$, (parallel to M_j in Sec. 3 but we replace j by c understood to be data dependent). Now if we treat the M_c as distinct models we have the alternative BMA

$$\hat{Y}^*_{BMA}(x) = \sum_{c=1}^{C} W(M_c \mid \mathcal{D}) \int_{[\alpha_{c-1},\alpha_c]} f(x \mid \theta) w_{[\alpha_{c-1},\alpha_c]}(\theta \mid \mathcal{D}) d\theta, \tag{74}$$

where

$$w_{[\alpha_{c-1},\alpha_c]}(\theta \mid \mathcal{D}) = \frac{w(\theta)p(y^n \mid \theta, x^n)\chi_{[\alpha_{c-1},\alpha_c]}}{\int_{[\alpha_{c-1},\alpha_c]} w(\theta)p(y^n \mid \theta, x^n)d\theta}.$$

Relative to the empirical BMA in (74), the alternative BMA in (74) is a 'psuedo'-BMA even though it can be regarded as a BMA in its own right. The key step now is to recognize the relationship between $1/B$ in (73) and the posterior weights in (74). Our analog to Prop. 1 for bagging with unidimensional θ is the following.

Proposition 4 *Assume the hypotheses of Lemma 1and Lemma 2. Then, for any x we have*

$$\left| \hat{f}_{bag}(x) - \hat{Y}^*_{BMA}(x) \right| \overset{P_T}{\to} 0.$$

Remark: In this proof we have used $\hat{F}$ and permitted the percentiles to be data dependent. If we used the population distribution (say F, the limit of the $\hat{F}$'s) of the $\hat{\theta}_b$'s instead of the empirical distribution, we would remove the dependence of the α_c on the data and the proof would go through similarly with the posterior probability accumulating in the one interval containing the true parameter value. We prefer $\hat{F}$ because it is data-driven.

Proof Write $G_n(\theta)$ for the posterior distribution function formed from $w(\cdot)$, the densities of Y_i, $p(y_i \mid \theta, x_i)$, and $\mathcal{D}_n$. Now define

$$A_c = \left\{ b \mid \hat{F}^{-1} G_n(\hat{\theta}_b) \in [\hat{F}^{-1} G_n(\alpha_{c-1}), \hat{F}^{-1} G_n(\alpha_c)] \right\} \tag{75}$$

for $c = 1, \ldots, C$, where $\hat{F}^{-1}$ exists because $\hat{F}$ is cadlag.

By Lemma 1, $\hat{\theta}$ is consistent so $\hat{\theta}_b$ is consistent and hence $\hat{F} \overset{P}{\to} \delta_{\theta_T}$, where $\delta_{\theta_T}(\theta)$ is 0 if $\theta < \theta_T$ and 1 otherwise. Also, by Lemma 2, $G_n \overset{P}{\to} \delta_{\theta_T}$. Thus, $\hat{F}^{-1}$ and G_n are inverses of each other asymptotically. Using this and the fact

$$\sum_{c=1}^{C} \#(A_c) = B, \tag{76}$$

i.e., the A_c's partition the B values $\hat{\theta}_b$, we have that by applying $\hat{F}$ to all three occurrences of $\hat{F}^{-1}$ in (75) we get

$$\frac{\#(A_c)}{B} \overset{P_T}{\to} \frac{1}{C}. \tag{77}$$

Now, define

$$\hat{\theta}_c = \frac{1}{\#(A_c)} \sum_{b \in A_c} \hat{\theta}_b. \tag{78}$$

Write (73) as

$$\hat{Y}_{bag}(x) = \sum_{c=1}^{C} \frac{\#(A_c)}{B} \left(\frac{1}{\#(A_c)} \sum_{b \in A_c} f(x \mid \hat{\theta}_b) \right), \tag{79}$$

and let

$$\hat{Y}_{app}(x) = \frac{1}{C} \sum_{c=1}^{C} f(x \mid \hat{\theta}_c) \quad \text{and} \quad \hat{Y}^*_{app}(x) = \int f(x \mid \theta) w(\theta \mid \mathcal{D}_n) d\theta. \tag{80}$$

By the triangle inequality,

$$\begin{aligned} \left| \hat{Y}_{bag}(x) - \hat{Y}^*_{BMA}(x) \right| \quad \leq \quad & \left| \hat{Y}_{bag}(x) - \hat{Y}_{app}(x) \right| & (81) \\ & + \left| \hat{Y}_{app}(x) - \hat{Y}^*_{app}(x) \right| & (82) \\ & + \left| \hat{Y}^*_{app}(x) - \hat{Y}^*_{BMA}(x) \right|. & (83) \end{aligned}$$

Now we want (81), (82), and (83) $\overset{P}{\to} 0$ to get Prop. 4. Indeed, for (81), observe that for all c, $\hat{\theta}_c \to \theta_T$ since $\hat{\theta}$ consistent. So, for all c and for all x,

$$\frac{1}{\#(A_c)} \sum_{b \in A_c} f(x \mid \hat{\theta}_b) - f(x \mid \hat{\theta}_c) \to f(x \mid \theta_T) - f(x \mid \theta_T) = 0. \tag{84}$$

This gives

$$\hat{Y}_{bag}(x) - \hat{Y}_{app}(x) = \sum_{c=1}^{C} \left(\frac{\#(A_c)}{B} - \frac{1}{C} \right) \left(\frac{1}{\#(A_c)} \sum_{b \in A_c} f(x \mid \hat{\theta}_b) \right)$$
$$+ \sum_{c=1}^{C} \frac{1}{C} \left(\frac{1}{\#(A_c)} \sum_{b \in A_c} f(x \mid \hat{\theta}_b) - f(x \mid \hat{\theta}_c) \right).$$

Using the boundedness of f and (77), the first term goes to 0. Using (84), the second term goes to 0. So, (81) $\overset{P_T}{\rightarrow}$ 0.

For (82),

$$\left| \hat{Y}_{app}(x) - \hat{Y}^*_{app}(x) \right| \quad \leq \quad \left| \frac{1}{C} \sum_{c=1}^{C} f(x \mid \hat{\theta}_c) - f(x \mid \theta_T) \right| \tag{85}$$

$$+ \left| f(x \mid \theta_T) - \int f(x \mid \theta) w(\theta \mid \mathcal{D}) d\theta \right|. \tag{86}$$

Since $\hat{\theta}_c \overset{P}{\rightarrow} \theta_T$ and $w(\theta \mid \mathcal{D}) \overset{P}{\rightarrow} \delta_{\theta_T}$, (85) and (86) $\rightarrow 0$ and hence (82) $\rightarrow 0$.

For (83) we have

$$\left| \hat{Y}^*_{app}(x) - \hat{Y}^*_{BMA}(x) \right| \leq \left| \int f(x \mid \theta) w(\theta \mid \mathcal{D}) d\theta - f(x \mid \theta_T) \right| \tag{87}$$

$$+ \left| f(x \mid \theta_T) - \sum_{c=1}^{C} W([\alpha_{c-1}, \alpha_c] \mid \mathcal{D}) \int_{[\alpha_{c-1}, \alpha_c]} f(x \mid \theta) w_{[\alpha_{c-1}, \alpha_c]}(\theta \mid \mathcal{D}) d\theta \right|. \tag{88}$$

(87) is the same as (86) so it goes to 0 whereas (88) is bounded above by

$$\sum_{c=1}^{C} W([\alpha_{c-1}, \alpha_c] \mid \mathcal{D}) \left| f(x \mid \theta_T) - \int_{[\alpha_{c-1}, \alpha_c]} f(x \mid \theta) w_{[\alpha_{c-1}, \alpha_c]}(\theta \mid \mathcal{D}) d\theta \right| \tag{89}$$

because the posterior probabilities are nonnegative and sum to 1. By definition $W([\alpha_{c-1}, \alpha_c] \mid \mathcal{D}_n)$ is bounded so we can ignore the first factor in the summands of (89). Under the consistency given by Lemma 1 we have that as $n, B \rightarrow \infty$, $\hat{\theta}_c \rightarrow \theta_T$. Also, by consistency, all percentiles tend to θ_T. That is, for $c = 1, \ldots, C$, $\alpha_c \rightarrow \theta_T$. Thus, by the Lebesgue differentiation theorem, as $C \rightarrow \infty$ (slowly) each $w_{[\alpha_{c-1}, \alpha_c]}(\theta \mid \mathcal{D})$ converges to unit mass at θ_T. (Actually, $w_{[\alpha_{c-1}, \alpha_c]}(\theta \mid \mathcal{D}_n)$ converges to whichever of α_{c-1} or α_c is closer to θ_T; when this is $c = 0$ so that $\alpha_0 = -\infty$, the one set $[\alpha_0, \alpha_1]$ contributes a constant times $1/C$ where $C \rightarrow \infty$.) So the second factors in the summands of (89) converge to zero, i.e., $\left| f(x \mid \theta_T) - \int_{[\alpha_{c-1}, \alpha_c]} f(x \mid \theta) w_{[\alpha_{c-1}, \alpha_c]}(\theta \mid \mathcal{D}) d\theta \right| \overset{P}{\rightarrow} 0$, and hence (89) $\overset{P}{\rightarrow}$ 0 which gives (83) $\overset{P}{\rightarrow}$ 0 pointwise in x. ■

The most severe limitation of this proposition is its restriction one unidimensional Θ. This can be removed by generalizing the definition in (75). We state this as a corollary.

Corollary 2 *Assume the hypotheses of Prop. 4 and that*

$$\sup_{i,x} E \mid Y_i(x) \mid^4 < \infty.$$

Then, we get the same limiting result as in Prop. 4 when θ ranges over $\Theta \subset \mathbb{R}^d$ satisfying $(\bar{\Theta})^0 = \Theta$ for any positive integer d.

Proof Let G_n be the posterior formed from $w(\theta)$, $p(y_i \mid \theta, x_i)$, and $\mathcal{D}_n$; Lemma 2 gives that G_n is asymptotically normal and identifies its asymptotic variance. Separately, let $\hat{F}$ be the empirical DF from the $\hat{\theta}_b$'s; effectively this is just the parametric bootstrap. Under the hypotheses of Lemma 1 and the extra moment condition, $\hat{F}$ is also asymptotically normal with the same mean and variance as G_n, see Rao (2017) for the statement and Hall (1992) Sec. 5.2 or van der Vaart (2020) Chap. 23 for details of proof. Without loss, we assume that $\hat{F}$ has been smoothed and so is no longer discrete.

Given that G_n and $\hat{F}$ have the same limiting distribution, we can generalize the definition of the $[\alpha_{c-1}, \alpha_c]$'s and the A_c's. To do this, we use the empirical marginals of $\hat{F}$ that we denote $\hat{F}_m$ for $m = 1, \ldots, d$. Let $(\alpha_1, \ldots, \alpha_d)$ be a vector of percentiles in which each α_m ranges over the values $\alpha_{m,c}$ where the $\alpha_{m,c}$'s are the c/C percentiles for $c = 1, \ldots C$ from $\hat{F}_m$. Now the dC vectors $(\alpha_{1,c_1}, \ldots, \alpha_{d,c_d})$ form a discrete set in d dimensions where c_m indicates the value of the percentile. If the spacing of the entries were regular, this would be a lattice, but we are using percentiles. Nevertheless, for ease of discussion, we call it a lattice. Each 'lattice' point defines a d-dimensional rectangle ('cuboid') by taking it as the geometric center with boundaries out to the boundaries of the nearest other vectors of length d. Denote the cuboids by $R_{c_1,\ldots,c_d}$. Now, as $n \to \infty$, $\hat{F}$, $P_{\hat{F}}(\theta_b \in R_{c_1,\ldots,c_d}) - G_n(R_{c_1,\ldots,c_d}) \to 0$ are both are bounded away from zero because they are based on percentiles.

Replacing the sets $[\alpha_{c-1}, \alpha_c]$ in (74) by $R_{c_1,\ldots,c_d}$ and the sets A_c in (75) by $A_{c_1,\ldots,c_d} = \{b \mid \hat{\theta}_b \in R_{c_1,\ldots,c_d}\}$ means we can apply the proof of Prop. 4 to give the result stated in Cor. 2. The only other difference is that probabilities like $1/C$ in (77) must be replaced by the empirical probabilities of the cuboids. ■

Given the result of Corollary 2 that shows how a bagging predictor and a pseudo-BMA are close, we present a result that represents the risk of a bagging predictor as asymptotically equivalent to a pseudo-BMA. It will be seen that the pseudo-BMA that splits up the parameter space into 'submodels' based on percentiles results in an expression that does not give a leading term. All terms in the pseudo-BMA are roughly equally important. However, by choosing the percentiles differently, a leading term can be identified asymptotically. The result is that bagging has the same limiting behavior as a pseudo-BMA. A different approach to understanding the MSE of bagging is taken in Sec. 6

Theorem 9 *Assume the hypotheses of Lemma 1and Lemma 2. Also, assume that $\Theta \subset \mathbb{R}^d$ is compact with $\overline{\Theta^0} = \Theta$ and that the explanatory variables range over a compact set.*

i) For any x_{n+1}

$$E_Y(Y(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 = \sigma^2 +$$
$$E_Y\Big(f(x_{n+1} \mid \theta_T) - W(R^* \mid \mathcal{D}) \textstyle\int_{R^*} f(x_{n+1} \mid \theta) w_{R^*}(\theta \mid \mathcal{D}) d\theta$$
$$- \textstyle\sum_{c \neq c^*} W(R_c \mid \mathcal{D}) \int_{R_c} f(x_{n+1} \mid \theta) w_{R_c}(\theta \mid \mathcal{D}) d\theta\Big)^2 + o(1),$$

where $R^* = R_{c^*} = R^*_{c_1^*,\ldots,c_d^*}$ *is the region in the parameter space (defined by the marginal percentiles) that contains* θ_T *and c is a vector of the form* $(c_1, \ldots, c_d)$*.*

ii) If $\theta_T \in \Theta^0$ *and we change from* C *uniformly spaced percentiles to using two percentiles corresponding to* $0 < \gamma_1 < \gamma_2 < 1$ *for each of the d dimensions of* θ*, then as* $\gamma_1 \to 0$ *and* $\gamma_2 \to 1$*, we get that*

$$E_Y(Y(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 = \sigma^2 +$$
$$E_Y\Big(f(x_{n+1} \mid \theta_T) - W(R^* \mid \mathcal{D}) \textstyle\int_{R^*} f(x_{n+1} \mid \theta) w_{R^*}(\theta \mid \mathcal{D}) d\theta\Big)^2 + o(1),$$

where $\theta_T \in R^*_{c_1,\ldots,c_d}$ *the region corresponding to the product of the mid-regions i.e., between* $[\gamma_1, \gamma_2]$*, for each dimension of* θ*.*

Remark: It is seen that this result is much like Theorem 5, even down to using a Bayesian approximation from Cor. 2.

Proof For the first clause of the theorem, we have

$$\begin{aligned}
&E_Y(Y(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 \\
&= E_Y(f(x_{n+1} \mid \theta_T) + \epsilon_{n+1} - \hat{Y}_{bag}(x_{n+1}))^2 \\
&= \sigma^2 + E_Y(f(x_{n+1} \mid \theta_T) - \hat{Y}_{bag}(x_{n+1}) \pm \hat{Y}^*_{BMA}(x_{n+1}))^2 \\
&= \sigma^2 + E_Y(f(x_{n+1} \mid \theta_T) - \hat{Y}^*_{BMA}(x_{n+1}))^2 + E_Y(\hat{Y}^*_{BMA}(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 \\
&\quad + 2E_Y\Big[(f(x_{n+1} \mid \theta_T) - \hat{Y}^*_{BMA}(x_{n+1}))(\hat{Y}^*_{BMA}(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))\Big]
\end{aligned}$$

By Cor. 2, $\Big|\hat{f}_{bag}(x_{n+1}) - \hat{Y}^*_{BMA}(x_{n+1})\Big| \xrightarrow{P_T} 0$. So, under the boundedness conditions here we have

$$E_Y(Y(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 = \sigma^2 + E_Y(f(x_{n+1} \mid \theta_T) - \hat{Y}^*_{BMA}(x_{n+1}))^2 + o(1),$$

which gives the first clause of the theorem.

The second clause of the Theorem follows by noting that since G_n and $\hat{F}$ have the same limiting normal $W(R^{*c}_{c_1,\ldots,c_d} \mid \mathcal{D}) \to 0$ as $n \to 0$ and the posterior probabilities multiply integrals that are bounded. ■

Theorem 9 shows that if we choose sets in the parameter space that are adaptive and equiprobable then all of them contribute roughly equally. So, c^* does not index a 'best' term. We only get an approximation of the MSE in terms of components; we do not find that the ensemble method is better than any of its components. On the other hand, if the percentiles are not equally spaced and converge to the tails of their respective marginal distributions then there is a leading term as identified in clause ii). It is similar to standard BMA studied in Sec. 3 in that a single term containing the correct model is the limit of the predictor. In either case, $E_Y(Y(x_{n+1}) - \hat{Y}_{bag}(x_{n+1}))^2 \to \sigma^2$ as expected when the model class contains a $\theta^*_{j^*}$ closest to the true model. Thus, bagging is asymptotically equivalent to a BMA using a data dependent prior and the model averaging properties of a bagged predictor do not seem to be representable in terms of true Bayesian model average.

This suggests that if we want our inferences to be extremely dependent on the data, as in bagging, and we want to use Bayesian methods, then we have to allow data dependent priors and these priors must either partition the posterior probability into roughly equal parts or simpy revert to a single dominating term. This reinforces Sec. 4 in Reid et al. (2003) where the authors note that strong matching of Bayes and Frequentist interval estimators is only possible for data dependent priors. The implication is that if an inference procedure is sufficiently strongly data-driven then only allowing the data to enter the posterior via the likelihood is not enough. The information in the data spills over to the prior. On the other hand, the way the data enters the prior is through the definition of subsets of the parameter space. Since the estimators for c^* or the α_c's are $\sqrt{n}$-consistent, they do not affect the convergence of the overall posterior. That is, introducing the parameters gives an empirical Bayes strategy that is equivalent asymptotically to a fully Bayes strategy.

6. Random Forests

Introduced by Breiman (2001), random forests (RF) are a modification of bagging trees based on the average of a large collection of de-correlated trees. The de-correlation, and hence variance reduction, is achieved in the tree-growing procedure via random selection of the input variables. Specifically, when growing a tree on a bootstrapped sample, before each split, select $m \leq p$ of the input variables at random as candidates for splitting. Values for m are typically $\sqrt{p}$ or $\log_2 p$. Random forests have essentially all the desirable convergence properties of bagging trees.

More formally, we modify the definitions of Breiman (2001) to our present context; the resemblance to the notation in Sec. 5 will be obvious. Consider a collection of regression functions $h(x, \Theta_k)$, for $k = 1, \ldots, K$ where the Θ_k are independent and identical copies of a random vector Θ with realized values $\theta_1, \ldots, \theta_K$. A random forest $\hat{Y}_{RF}(x)$ is any stochastically symmetric function of the $h(x, \Theta_k)$'s as functions of x, i.e., each randomly chosen $h(x, \Theta_k)$ has the same influence on the predicted value of $Y(x)$. RF's of trees for classification satisfy this definition, but many other formulations are possible. Here we focus on regression and we take means of the $h(x, \theta_k)$'s over k. This definition of RF's includes bagging as used in Sec. 5 and the original formulation of RF's (random selection of variables for splits in trees) as a special cases.

Write $\hat{Y}_{RF}(\cdot)$ for the random forest point predictor, parallel to (73), for some base function $f(x \mid \theta)$. We start by observing that the results from the previous subsection for $\hat{Y}_{bag}$ continue to hold for $\hat{Y}_{RF}$. Then we give sufficient conditions for a result in Breiman (2001) to hold thereby giving a more satisfactory analog to Theorems 2, 5, 8, for RF's than Theorem 9 is. We begin with the following.

Theorem 10 *Assume a signal-plus-noise model of the form* (72) *and assume the hypotheses of Lemma 1 and Lemma 2. Then, Prop. 4, Cor. 2, and Theorem 9 continue to hold with* $\hat{Y}_{bag} = \hat{f}_{bag}$ *replaced by* $\hat{Y}_{RF}$*. That is,* $\hat{Y}_{RF}$ *can be approximated asymptotically by a pseudo-BMA of the form* (74) *and satisfies the two MSE bounds given in Theorem* (9)*.*

Proof The proofs of these statement follow from reading the proofs of Prop. 4, Cor. 2, and Theorem 9 and verifying that each step follows if the RF procedure is used (random selection of, say, $\sqrt{p}$, the explanatory variables) in place of the bagging procedure. ■

This result gives versions of the first and third sorts of results we want for model averages. The second sort of result—an oracle inequality or bounds on empirical risks—remains too hard to obtain in the generality sought here.

Since the third sort of result—verifying that the model average is a better predictor in some sense than its components—is the most important of the three, we recall Theorem 11.2 in Breiman (2001). It is stated for trees, but in fact generalizes to applying the random forest procedure to any rich and well-enough behaved class of predictors. It states, informally, for RF regression (with trees), that if an unbiasedness condition (stated in Clause ii)) is satisfied then the generalization error of the RF is bounded by a weighted correlation times the generalization error of a single tree. The correlation is between two random trees such as might be in the forest. That is, a multiple of the MSE of a random tree bounds the MSE of the RF. Note that the decorrelation achieved by the random selection of explanatory variables lowers the weighted correlation to give a tighter bound on the MSE of the RF.

Theorem 11.2 is qualitatively different from the bounds on MSE from earlier results here such as Theorems 2, 5, 8, 9, and 10 that focus on determining a leading term. Here, as in Sec. 5, there is no 'leading' term; all terms contribute equally, if only in a probabilistic sense. Indeed, Clause ii) treats all the terms in the RF as if they are equivalent and obtains a bound in terms of their generic form. This equivalence of terms can also be seen in Clause i) of Theorem 9 for bagging, and the discussion after it and recurs in Theorem 10.

Specifically, Breiman (2001) shows the following in general, i.e., no signal-plus-noise model need be assumed.

Proposition 5 *(***Breiman 2001** *Sec. 11) Clause i): Assume that* Y *has bounded second moment and that* $h(x, \theta_k)$ *is continuous on an open domain* $\mathcal{X} \times \Theta\mathbb{R}^{p+d}$ *that satisfies* $(\overline{\mathcal{X} \times \Theta})^0 = \mathcal{X} \times \Theta$*. Then,*

$$E_{X,Y}(Y - \overline{h(X, \Theta_k)})^2 \xrightarrow{a.s.} E_{XY}(Y - E_\Theta h(X, \Theta))^2$$

Clause ii) Assume that for all Θ*,* $E(Y) = E_X h(X, \Theta)$*. Then,*

$$E_{XY}(Y - E_\Theta h(X, \Theta))^2 \leq \overline{\rho}\ E_\Theta E_{X,Y}(Y - h(X, \Theta))^2,$$

where $\overline{\rho}$ *is the weighted correlation between any pair of residuals* $Y - h(X, \Theta)$ *and* $Y - h(X, \Theta')$ *for independent copies* Θ *and* Θ' *of* Θ*.*

The proof of Clause i) is simply the strong law of large numbers applied to outcomes of Θ. The proof of Clause ii) follows from manipulating the left hand side to reveal the correlation. The hypothesis of Clause ii) comes from recognizing correlations involve division by standard deviations that themselves must be expectations of centered random variables. Both clauses apply for RF's and bagging.

What does the unbiasedness condition $E(Y) = E_X h(X, \Theta)$ mean? First, it assumes that Θ is a random variable with outcomes $\Theta = \theta$. Second, both Y and X are random. This is a common assumption in classification which Prop. 5 uses in regression. In fact, the goal is often phrased as minimizing $E(Y - h(X))^2$ over functions h meaning that X and Y are random and no specific assumptions are made about any relationship between them. Essentially, this is the $\mathcal{M}$-complete case described in Bernardo and Smith (2000) or Clarke and Clarke (2018).

The key issue is satisfying the unbiasedness condition so that Clause ii) can be used. The intuition behind unbiasedness is that for any $\Theta = \theta$, the function $h(x, \theta)$ when averaged over X gives $E(Y)$. That is, in the distrubution F_X, $h(X, \theta)$ neither overestimates Y nor underestimates Y. Let $h_{\mathsf{init}} \in \mathcal{H}$, a class of functions that forms a Banach space that has a countable dense subset. Loosely, given F_X (and F_Y), if $\mathcal{H}$ is large enough then the unbiasedness condition can be satisfied. Without loss of generality assume $E(Y) < E_X h_{\mathsf{init}}(X, \theta)$ for some $\Theta = \theta$. We can construct for θ a function $h(x, \theta)$ that satisfies $E(Y) = E_X h(X, \theta)$. Let $B(f(x, \theta), R)$ be a ball of radius R in $\mathcal{H}$ centered at $f(x, \theta)$, where R is chosen large enough that for some $h_{\mathsf{end}} \in B(f(x, \theta), R)$, $E(Y) > E_X h_{\mathsf{end}}(X, \theta)$. Consider the line in $\mathcal{H}$ defined by $\gamma h_{\mathsf{init}}(x, \theta) + (1-\gamma) h_{\mathsf{end}}(x, \theta)$. Now, since a dense set exists, for some $\gamma \in (0, 1)$ there must be a γ^* so that $h(\cdot, \theta) = \gamma^* h_{\mathsf{init}}(x, \theta) + (1 - \gamma^*) h_{\mathsf{end}}(x, \theta) \in \mathcal{H}$ with $E(Y) = E_X h(X, \theta)$, as required. Thus, given $h_{\mathsf{init}}(x, \theta)$ we can find an $h(x, \theta)$ that satisfies the unbiasedness condition on the line joining two elements of $\mathcal{H}$. This simple argument does not by itself ensure that $h(x, \theta_1)$ and $h(x, \theta_2)$ will be distinct when $\theta_1 \neq \theta_2$. However, it is easy to see that rich enough classes $\mathcal{H}$, such as trees and neural nets, will usually have distinct elements. We summarize this reasoning in the following.

Proposition 6 *For given θ, let $h_{\mathsf{init}}(x, \theta) \in \mathcal{H}$, a Banach space with a countable dense set. Then:*

i) $\exists h(x, \theta) \in \mathcal{H}$, derived from h_{init}, so that $E(Y) = E_X h(X, \theta)$, and,

ii) As long as $h_{\mathsf{init}}(x, \theta_1)$ and $h_{\mathsf{init}}(x, \theta_2)$ are distinct a.e. and they are joined by a straight line to functions $h_{\mathsf{end}}(x, \theta_1)$ and $h_{\mathsf{end}}(x, \theta_2)$ that are also distinct a.e., then $h(x, \theta_1)$ will be distinct from $h(x, \theta_2)$ a.e.

Taken together, Props. 5 and 6 give a version of the third kind of result – a sense in which an ensemble method performs better than any of its components.

7. Boosting

A sixth model averaging technique is boosting, see Schapire (1990), one of the most successful procedures for classification. While there are several boosting algorithms, AdaBoost due to Freund and Schapire (1996) is arguably the most popular. Boosting can also be extended to regression problems, see Freund and Schapire (1997) and Ridgeway et al. (1999b), amongst others. The basic idea behind the extension of boosting for classification to boosting for regression is to discretize the response Y and apply boosting to each interval range separately. Effectively, this generates an approximation to Y. Boosting for regression has not received as much attention as boosting for classification; see, for instance, Schapire and Freund (2012). Nevertheless, we argue that boosted regression, like other ensembling methods, is consistent and performs better than any of its components. As a final point for this section we give a Bayesian interpretation for boosted regression.

The concept of boosting in regression requires a different problem formulation than in earlier sections. Regard the pairs $(x_i, y_i(x_i))$ for $i = 1, \ldots, n$ as IID outcomes from a bivariate probability P on $\mathcal{X} \times \mathcal{Y}$ for which $\mathcal{X} = \overline{\mathcal{X}^0} \subset \mathbb{R}^p$ and $\mathcal{Y} = [0,1]$. We seek a function $h : \mathcal{X} \to \mathcal{Y}$ for which $d(Y, h(X))$ is satisfactorily small for some distance d. Here we use expected squared error in P. This differs from Sec. 6 because here we have $Y = Y(x)$. The effect is that we will get optimality in an MSE sense instead of the optimality we would get in a signal-plus-noise setting where optimal predictors generally achieve σ^2 as their asymptotically minimum variance. It is left as an exercise to specialize the results below to a signal-plus-noise model and obtain asymptotics similar to what we showed in earlier sections. This means that we are including $\mathcal{M}$-complete problems.

Following Freund and Schapire (1997) Sec. 5.3, let $\hat{Y}_{n,T}(x)$ be the boosted regression function after T iterations and let $\hat{Y}(\cdot) = \arg\min_h E_P(Y - h(X))^2$. By construction $\hat{Y}_{n,T}$ converges a.e. to $\hat{Y}_n = \arg\min_h E_{emp}(Y - h(X))^2$, where E_{emp} is the expectation under the empirical distribution function. This follows from applying the monotone convergence theorem to the inequality in Theorem 12 of Freund and Schapire (1997), as $T \to \infty$, in the empirical probability for (X, Y).

As in previous sections, we start with a consistency result. To see that $\hat{Y}_n(\cdot) \xrightarrow{P} \hat{Y}(\cdot)$ as $n \to \infty$ we use the following generalization of the Newey-McFadden theorem from finite dimensional parameter spaces to function spaces.

Proposition 7 *(Generalization of Theorem 2.1 in Newey and McFadden (2012)) Suppose $\hat{Y}_n$ achieves $\min_h E_{emp}(Y - h(X))^2$ and $\hat{Y}$ achieves $\min_h E_P(Y - h(X))^2$. Also, assume*

(i) $E_{emp}(Y - h(X))^2 \xrightarrow{P} E_P(Y - h(X))^2$ uniformly for $h \in K$, a compact set in a Hilbert space $\mathcal{H}$, under the weak topology,* [1]

(ii) h uniquely achieves $\min_h E_P(Y - h(X))^2$,

(iii) $E_P(Y - h(X))^2$ is continuous as a real valued function of h in the weak topology.*

Then, $\hat{Y}_n \xrightarrow{P} \hat{Y}$.

Proof The proof is the same as in Newey and McFadden (2012) but in the weak* topology. Recall, that the dual space to $\mathcal{H}$ is $\mathcal{H}^*$, also a Hilbert space, and $\mathcal{H}^*$ is isomorphic to $\mathcal{H}$. Hence $\mathcal{H}^{**}$, the dual of $\mathcal{H}^*$, is a Hilbert space isomorphic to $\mathcal{H}$. The Alaoglu theorem states that the closed unit ball in $\mathcal{H}$ is compact under the weak* topology and every bounded closed set in $\mathcal{H}$ is relatively weakly compact in the weak* topology. All the steps in the

[1] In a Hilbert space, the Reisz representation theorem shows that $\mathcal{H}^*$ is isomorphic to its dual $\mathcal{H}$. So, for Hilbert spaces, the weak and weak* topologies coincide and are metrizable. We retain the distinction between the weak and weak* topologies because they are not in general isomorphic for Banach spaces. We hope to generalize our results to functions in a Banach space since the inner product on $\mathcal{H}$ is not explicitly used in the proof.

proof of the Newey-McFadden theorem can now be done in $\mathcal{H}$ with the weak* topology: Simply replace steps using finite dimensional real spaces under the usual real topology with the corresponding steps using $\mathcal{H}$ under the weak* topology. ■

This generalized Newey-McFadden theorem now gives $\hat{Y}_n = \arg\min_h E_{emp}(Y - h(X))^2 \xrightarrow{P} \hat{Y} = \arg\min_h E_P(Y - h(X))^2$ as $n \to \infty$. Taken altogether, we have the following consistency theorem.

Theorem 11 *Suppose $n \to \infty$ and $T = T_n \to \infty$ at a suitable rate in n. Then, we have that $\hat{Y}_{n,T}(x) \xrightarrow{P} \hat{Y}(x)$.*

Not that this is consistency of $\hat{Y}_{n,T}(x)$ for $\hat{Y}(x)$ and that by construction $\hat{Y}(x)$ is consistent for $\min_h E_P(Y - h(X))^2$ automatically giving the following corollary.

Corollary 3 *Under the hypotheses of Theorem 11, $\hat{Y}_{n,T}(x)$ is consistent for $\min_h E_P(Y - h(X))^2$.*

As in the previous two sections, obtaining an oracle inequality or other bounds on the empirical risk is very difficult. One starting point is Freund and Schapire (1997) Theorem 7 but it is for an iteration not cumulative over many instances. This is left as a gap in the characterization of boosting techniques for regression (and classification).

Turning to the third sort of result we want, Prop. 7 also gives that the boosted regression function is better than any of its terms, or indeed any selection of finitely many of its terms, unless the final selection of terms is asymptotically equivalent to the boosted regression function. Thus, we have a result to parallel our earlier results comparing model averages with their components. Indeed, Cor. 4 below is suggested by the fact that $\hat{Y}_{n,T}$ from Freund and Schapire (1997) p. 136 col. 1 optimizes an empirical squared error.

Corollary 4 *Let $\hat{Y}_{n,\ell}$ be a regression function formed from a finite collection of terms in $\hat{Y}_n(\cdot)$. Then,*

$$\liminf_{n\to\infty} \left[E_P(\hat{Y}_{n,\ell} - h(X))^2 - E_P(\hat{Y}_n - h(X))^2 \right] \geq 0$$

Proof Obvious because $\hat{Y}_n$ asymptotically minimizes $E_P(Y - h(X))^2$. ■

While this result concludes the treatment of the three sorts of results we want for model average predictors, we note that earlier model averages bore a close resemblance to Bayesian predictors. Indeed, BMA is Bayes, bagging and random forests were shown to be approximately Bayes, and stacking has been shown to be a Bayes action; see Le and Clarke (2017) under several loss functions. So, to conclude this section we turn to giving a Bayesian interpretation to boosted regression functions. We do this by relating boosted regression to boosted classifiers. Our technique is modeled on Ridgeway et al. (1999a). Given h as above, define $\tilde{h} : \{x\} \times \{y\} \to \{0, 1\}$ where h and $\tilde{h}$ are related by $\tilde{h}(x, y) = 1_{\{y \geq h(x)\}}$. Given an $\tilde{h}$ we can also recover an h. Suppose $\mathcal{X}$ is the set of x-values and $S_m = \{0, 1/m, \ldots, (m-1)/m, 1\}$.

If $\mathcal{X}$ has the n x-values from $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \ldots, n\}$ then $\#(\mathcal{X} \times S_m) = mn$ and we can consider the set

$$\mathcal{D}^* = \{(x, s, y^*) \mid x \in \mathcal{X}, s \in S_m, y^* = 1_{\{s \geq y(x)\}}\}.$$

Since $y^* \in \{0, 1\}$, $\mathcal{D}^*$ is a (binary) classification data set derived from $\mathcal{D}$. Now, if $\tilde{h}(x, s)$ is a classifier for $\mathcal{D}^*$, i.e., $\tilde{h}(x, s)$ predicts the class of $y^* = y^*(x, s)$, we can define the regression function

$$h_m(x) = \inf_{s \in S_m} \{s \mid \tilde{h}(x, s) = 1\}. \tag{90}$$

Intuitively, passing from $\mathcal{D}$ to $\mathcal{D}^*$ converts a regression problem to a classification problem by assigning to each x the pair of intervals $\{s < y(x)\}$ and $\{s \geq y(x)\}$ that are coded as $0, 1$, respectively, to a fineness of $1/m$. Hence, $h_m(x)$ gives the breakpoint in y-space as the prediction for $Y(x)$.

We can now apply the AdaBoost algorithm from Freund and Schapire (1997) to $\mathcal{D}^*$. That is, let $\hat{C}_t(x, s)$ denote the iterates for $t = 1, \ldots, T$ from AdaBoost and form the classifier

$$BSTC_T(x, s) = \mathsf{sign}\left(\sum_{t=1}^{T} \hat{\beta}_t \hat{C}_t(x, s)\right) \tag{91}$$

where the weights $\hat{\beta}_t$ are also given in the AdaBoost algorithm. Then $BSTC_T$ plays the role of $\tilde{h}$.

It was shown in Le and Clarke (2018) that $BSTC_T$ converges to the Bayes classifier and hence is Bayes optimal, asymptotically, as $T \to \infty$. More formally,

$$\sum_{t=1}^{T} \hat{\beta}_t \hat{C}_t(x, s) \to \log \frac{P(Y = 1 \mid X = x, S = s)}{P(Y = 0 \mid X = x, S = s)}. \tag{92}$$

Hence, for fixed x and m, using (91) in (90) gives a regression function that inherits optimality from the asymptotically Bayes classifier. The size of m matters here because if m is large enough then each distinct point in $\mathcal{D}$ corresponds to a well defined set of points in $\mathcal{D}^*$, i.e., the inclusion map $\mathcal{D} \to \mathcal{D}^*$ is one-to-one.

Assuming limits are continuous and convergence is uniform, (92) gives

$$\sum_{t=1}^{T} \hat{\beta}_t \hat{C}_t(x, z) \to \log \frac{P(Y = 1 \mid X = x, z)}{P(Y = 0 \mid X = x, z)}$$

when $s = s_k \to z$ independently of the convergences with T or $n \to \infty$. Thus we have $BSTC(x, s_k) \to BSTC(x, z)$ as m and $k \to \infty$, i.e., the limit on the right hand side exists. Now, we can set

$$h(x) = \inf_z \{z \mid BSTC(x, z) = 1\} \tag{93}$$

and it is the limit of $\hat{Y}_{n,T}$ and $\hat{Y}_n$. Indeed, (93) is essentially the same as the output from Algorithm 5 in Freund and Schapire (1997). That is, the boosted regression function comes from the boosted classifier that is asymptotically Bayes. This provides a Bayes interpretation for boosted regression.

A readily formalizable argument reinforces this parallel to Cor. 4. Let $\hat{Y}^*$ be any regression function for Y based on $\mathcal{D}$ other than the boosted regression function $\hat{Y}_m$ where $\hat{Y}_m(x) = \inf_{s \in S_m}\{s \mid \tilde{h}(x,s) = 1\}$. Then, for any x,

$$|\hat{Y}_m(x) - \hat{Y}^*(x)| \leq |\hat{Y}_m(x) - \hat{Y}_m(x_c)| + |\hat{Y}_m(x_c) - \hat{Y}^*(x_c)| + |\hat{Y}^*(x_c) - \hat{Y}^*(x)| \tag{94}$$

where $x_c = \arg\min_k |x - x_k|$. Since $(x_i, y_i)|_{i=1}^n$ becomes dense in $[0,1] \times [0,1]$, $x_c - x \overset{P}{\to} 0$. So, provided $\hat{Y}_m(x)$ and $\hat{Y}^*(x)$ are bounded as functions of x, the first and third terms of (94) $\overset{P}{\to} 0$. Now, since $\hat{Y}_m(x_c) \to y(x_c)$ as $m \to \infty$ and $\hat{Y}_m(x_c)$ has tolerance $1/m$ around $y(x_c)$, if $\hat{Y}_m(x_c) \geq \hat{Y}^*(x_c)$ the second term in (94) can be bounded by

$$0 \leq \hat{Y}_m(x_c) - \hat{Y}^*(x_c) \leq (y(x_c) - \hat{Y}^*(x_c)) + \frac{1}{m}. \tag{95}$$

Otherwise, if $\hat{Y}_m(x_c) < \hat{Y}^*(x_c)$ the second term in (94) can be bounded by

$$0 < \hat{Y}^*(x_c) - \hat{Y}_m(x_c) \leq (\hat{Y}^*(x_c) - y(x_c)) - \frac{1}{m}. \tag{96}$$

In any event, (95) or (96) means if $y(x)$ is continuous, the closer $\hat{Y}^*(x)$ is to $y(x)$ the better $\hat{Y}^*(x)$ is and any regression function not close enough to $\hat{Y}_m$ would be outperformed by $\hat{Y}_m(\cdot)$. So regression functions too different from the boosted regression will be worse than $\hat{Y}_m$ even if they are made from terms in $\hat{Y}_m$ unless they are equivalent to $\hat{Y}_m$.

8. Concluding Remarks

First, good predictors that are model averages tend to outperform any of their components. Second, good predictors that are not Bayes are nearly Bayes. This is seen in our results that show that bagging and RF's are pseudo-BMA's in a limiting sense, that boosting for regression has a Bayes justification (from classification), and that stacking is asymptotically a Bayes action. Moreover, MMA is based on regression and it is well-known that Bayesian linear regression under normal priors is equivalent to regular frequentist regression as the priors become more spread out. So, it is reasonable to conjecture that MMA is is Bayesian analog will be asymptotically identical. While our other sorts of results—consistency and bounds on empirical risk—are important for understanding and using model average predictors, they do not justify why model averages are generally better than other predictors.

The biggest problem left insufficiently addressed here is the selection of the model list $\mathcal{F}$. However, design principles for $\mathcal{F}$ have been examined by numerous other authors. First, Breiman (1996a) argued that the f_j's should be as different as possible from each other. Second, George (1999) identified a phenomenon called dilution in which the models on a model list close to f_T are so numerous that each has very small probability leading to BMA predictors that are essentially always zero. The same phenomenon can occur with some frequentist predictors as well. In the Bayes context, Chipman et al. (2001) proposed prior selection techniques to avoid dilution. Third, Occam's window, Madigan and Raftery (1994), is another model list selection principle. It's the opposite of dilution and is well suited to prediction because a dynamic form of it has recently been proposed, see Onorante and Raftery (2014). Fourth, Yu et al. (2013) argues that choosing different models on different parts of the covariate space provides better prediction.

Finally, fifth, if some data is held in reserve for model list selection, variance-bias tradeoff arguments may help guide the selection of $\mathcal{F}$. In MSE terms, small, localized $\mathcal{F}$'s will tend to have high bias, while large spread out $\mathcal{F}$'s will tend to have high variance. This suggests $\mathcal{F}$ should be chosen so that it is believed that f_T is in the 'middle' of $\mathcal{F}$ rather than on its periphery or outside its closed convex hull in order to outperform well-selected individual models. This is consistent with Clarke and Fokoue (2011) that argued one can find MSE optimal model lists. Of course, in the case that there is no true model it is possible for model selection and downstream prediction to oscillate within a collection of predictors, some of which may be mixtures of models. Hence, optimal model list selection must still be regarded as an incompletely resolved issue for model averaging techniques just as it is an incompletely resolved issue for model selection.

Appendix A. Supporting Results for Sec. 2

In this Appendix we present four Lemmata. Two are results for asymptotic normality of the MLE and posterior when the observations are independent but not identical. In standard cases, these are well known; see Hoadley (1971) and Hartigan (1983) for instance. However, here the non-identicality is due to an explanatory variable not just the parametric family. The second two results are analogs of these results when none of the models can be assumed correct. These latter two Lemmata generalize the first two and generalize White (1982) and Berk (1966). Note that in this section, j indexes model classes rather then the entries in a vector defining an element of a model class.

A.1 INID results – correct model

We begin by establishing asymptotic Normality of MLE for independent but not identical variables. Our proof is modified from the proof of the analogous result in Theorem 18 of Ferguson (1996). Consider the following list of hypotheses.

(i) Let $\theta_j \in \Theta_j$, an open subset of $\mathbb{R}^{d_j}$ such that $(\bar{\Theta}_j)^o = \Theta_j$, i.e., Θ_j is the interior of its closure.

(ii) All second partial derivatives of $p_j(y|x, \theta_j)$ w.r.t. θ_j exist and are continuous for all x, y, and may be passed under the integral sign in $\int p_j(y|x, \theta_j)dy$.

(iii) There exists a function $K(\cdot)$ such that $E_{\theta_j} K(Y) < \infty$ and each element of the matrix $\left(\frac{\partial^2}{\partial\theta_\ell \partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right)_{\ell,k=1,\ldots,d_j}$ is bounded in absolute value by $K(y)$ uniformly for θ_j in an open neighborhood of $\theta_{0,j}$.

(iv) Assume that, for any x_i,

$$I_i(\theta_{0,j} \mid x_i) = \left(-E_{\theta_{0,j}} \left[\frac{\partial^2}{\partial\theta_\ell \partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right]\right)_{\ell,k=1,\ldots d_j}$$

is continuous on an open set around the true value $\theta_{0,j}$ and is positive definite at $\theta_{0,j}$.

(v) For any $\theta_{j,0} \in \Theta_j$ and any x_i, $p_j(y \mid x_i, \theta_j) = p_j(y \mid x_i, \theta_{j,0})$ a.e. in y implies $\theta_j = \theta_{j,0}$.

(vi) The average $(1/n)\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i)$ is invertible for any n and the $d_j = \dim(\theta_j)$-vector

$$\Psi(\theta_{j,0} \mid x, y) = \left(\nabla_{\theta_j} \log p_j(y \mid x, \theta_{j,0})\right)^T$$

satisfies

$$\sum_{i=1}^n E\left\| \frac{1}{n}\Big[\sum_{k=1}^n I_k(\theta_{0,j} \mid x_k)\Big]^{-1/2} \Psi(\theta_{j,0} \mid x_i, Y)\right\|^3 = o\left(\frac{1}{n^{3/2}}\right).$$

Note that assumptions (i) to (v) are for the existence of $\hat{\theta}_{n,j}$ analogous to those for Cramer's Theorem in the IID case. Assumption (vi) allows us to identify the variance in the asymptotic normal distribution. We have the following.

Lemma 1 *Let $y_i \sim p_j(y \mid x_i, \theta_j)$ be INID and let $\theta_{j,0}$ denote the true value of the parameter. Assume hypotheses (i) – (vi). If $\hat{\theta}_{n,j}$ is the MLE of $\theta_{j,0}$ then*

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_{j,0}) - N\left(0, \left[\frac{1}{n}\sum_{i=1}^n I_i(\theta_{j,0} \mid x_i)\right]^{-1}\right) \xrightarrow{L} 0.$$

Proof For ease of exposition, we write $\dot{\Psi}(\theta_j \mid x, y) = \left(\nabla^2_{\theta_j} \log p_j(y \mid x_i, \theta_j)\right)$ for the $d_j \times d_j$ matrix of second partial derivatives of $\log p_j(y \mid x, \theta_{j,0})$. So, $I(\theta_j \mid x) = -E_{\theta_j}\dot{\Psi}(\theta_j \mid x, Y)$. The likelihood is $L_n(\theta_j \mid \mathcal{D}_n) = \prod_{i=1}^n p_j(y_i \mid x_i, \theta_j)$ with derivative of its logarithm given by the d_j-vector $\dot{\ell}_n(\theta_j \mid \mathcal{D}_n) = \nabla_{\theta_j} \log L_n(\theta_j \mid \mathcal{D}_n) = \sum_{i=1}^n \nabla_{\theta_j} \log p_j(y_i \mid x_i, \theta_j)$.

Now, the Mean Value Theorem gives

$$\dot{\ell}_n(\theta_j \mid \mathcal{D}_n) = \dot{\ell}_n(\theta_{0,j} \mid \mathcal{D}_n) + \int_0^1 \sum_{i=1}^n \dot{\Psi}(\theta_{0,j} + \lambda(\theta_j - \theta_{0,j}) \mid x_i, y_i)\mathrm{d}\lambda(\theta_j - \theta_{0,j}).$$

Letting $\theta_j = \hat{\theta}_{n,j}$ and dividing by $\sqrt{n}$ gives

$$\frac{1}{\sqrt{n}}\dot{\ell}_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n) = B_n\sqrt{n}(\hat{\theta}_{n,j} - \theta_{0,j}), \tag{97}$$

where

$$B_n = -\int_0^1 \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(\theta_{0,j} + \lambda(\hat{\theta}_{n,j} - \theta_{0,j}) \mid x_i, y_i)d\lambda.$$

Observe that $E_{\theta_{0,j}}\Psi(\theta_{0,j} \mid x, Y) = 0$ and $Var_{\theta_{0,j}}\Psi(\theta_{0,j} \mid x, Y) = I(\theta_{0,j} \mid x)$. So, from the Berry-Esseen Theorem for independent but non-identical multivariate random variables, as $n \to \infty$, Assumption (vi) gives the weak convergence meant by

$$\frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_j \mid \mathcal{D}_n) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \Psi(\theta_{0,j} \mid x_i, y_i)\right) \overset{L}{\approx} N\left(0, \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i)\right). \tag{98}$$

It remains to show $B_n - (1/n)\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \xrightarrow{P} 0$. Indeed, let $\epsilon > 0$ and note that from Assumptions (ii) and (iii) $E_{\theta_j}\dot{\Psi}(\theta_j \mid x, Y)$ is continuous in θ_j and x. Thus, there is a $\rho > 0$ such that $|\theta_j - \theta_{0,j}| < \rho$ implies

$$\sup_{x \in \mathcal{X}} |E_{\theta_j}\dot{\Psi}(\theta_j \mid x, Y) + I(\theta_{0,j} \mid x)| < \epsilon. \tag{99}$$

Also, Assumptions (ii) and (iii) give that with probability one there is an integer N such that if $n > N$ then

$$\sup_{|\theta_j - \theta_{0,j}| < \rho; x_i \in \mathcal{X}} \left| \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(\theta_j \mid x_i, y_i) - \frac{1}{n}\sum_{i=1}^n E_{\theta_j}\dot{\Psi}(\theta_j \mid x_i, Y) \right| < \epsilon. \tag{100}$$

By consistency, there is an N is so large that for $n > N$, $|\hat{\theta}_{n,j} - \theta_{0,j}| < \rho$ with probability at least $1 - \eta$ for some small $\eta > 0$ (under $\theta_{0,j}$).

Now, we have that

$$\begin{aligned}
&\left| B_n - \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \right| \\
&\le \int_0^1 \left| \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(\theta_{0,j} + \lambda(\hat{\theta}_{n,j} - \theta_{0,j}) \mid x_i, y_i) + \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \right| d\lambda \\
&\le \int_0^1 \left| \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(\theta_{0,j} + \lambda(\hat{\theta}_{n,j} - \theta_{0,j}) \mid x_i, y_i) - \frac{1}{n}\sum_{i=1}^n E_{\theta_j}\dot{\Psi}(\theta_j \mid x_i, Y) \right| \\
&\qquad + \left| \frac{1}{n}\sum_{i=1}^n E_{\theta_j}\dot{\Psi}(\theta_j \mid x_i, Y) + \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \right| d\lambda \\
&\le \int_0^1 \sup_{\theta_j : |\theta_j - \theta_{0,j}| < \rho; x \in \mathcal{X}} \left| \frac{1}{n}\sum_{i=1}^n \dot{\Psi}(\theta_j \mid x, y_i) - \frac{1}{n}\sum_{i=1}^n E_{\theta_j}\dot{\Psi}(\theta_j \mid x, Y) \right| \\
&\qquad + \sup_{x_i \in \mathcal{X}} \frac{1}{n}\sum_{i=1}^n \left| E_{\theta_j}\dot{\Psi}(\theta_j \mid x_i, Y) + \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \right| d\lambda \\
&\le \epsilon + \frac{1}{n}\sum_{i=1}^n \epsilon = 2\epsilon \qquad \text{by (99) and (100).}
\end{aligned}$$

This gives that

$$B_n - \frac{1}{n}\sum_{i=1}^n I_i(\theta_{0,j} \mid x_i) \xrightarrow{P} 0. \tag{101}$$

Therefore, from (97), (98), and (101), Slutsky's Theorem gives the conclusion of Lemma 1. ■

Having established asymptotic normality of the MLE under INID conditions, we turn to asymptotic normality of the posterior. This time our proof is modeled on Theorem 21 in Ferguson (1996).

Lemma 2 *Assume hypotheses (i) – (vi) and that the prior $w(\theta_j) > 0$ is continuous for all $\theta_j \in \Theta$. Then the posterior $w_j(\theta_j \mid \mathcal{D}_n)$ is asymptotically normal, i.e.,*

$$w_j(\theta_j \mid \mathcal{D}_n) \overset{a.s.}{\approx} N\left(\hat{\theta}_{n,j}, \frac{1}{n}\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} I_i(\theta_{0,j} \mid x_i)\right]^{-1}\right).$$

Proof We continue using the notation defined in the proof of Lemma 1. Write

$$\begin{aligned}\frac{L_n(\theta_j \mid \mathcal{D}_n)}{L_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)} &= \exp\{\ell_n(\theta_j \mid \mathcal{D}_n) - \ell_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)\} \\ &= \exp\{-n(\theta_j - \hat{\theta}_{n,j})^T A_n(\theta_j \mid \mathcal{D}_n)(\theta_j - \hat{\theta}_{n,j})\}, \qquad (102)\end{aligned}$$

where we have Taylor expanded $\ell_n(\theta_j \mid \mathcal{D}_n) = \log L_n(\theta_j \mid \mathcal{D}_n)$ about $\hat{\theta}_{n,j}$ to get $\ell_n(\theta_j \mid \mathcal{D}_n) - \ell_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)$ equals

$$\dot{\ell}_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)(\theta_j - \hat{\theta}_{n,j}) - n(\theta_j - \hat{\theta}_{n,j})^T A_n(\theta_j \mid \mathcal{D}_n)(\theta_j - \hat{\theta}_{n,j})$$

in which

$$A_n(\theta_j \mid \mathcal{D}_n) = -\frac{1}{n}\int_0^1\int_0^1 v\ddot{\ell}_n(\hat{\theta}_{n,j} + uv(\theta_j - \hat{\theta}_{n,j}))dudv \qquad (103)$$

and we have the fact that $\dot{\ell}_n(\hat{\theta}_{n,j}) = 0$ w.p.1.

Let $\eta_j = \sqrt{n}(\theta_j - \hat{\theta}_{n,j})$. Using (99) and (100) we can rewrite (103) as

$$\begin{aligned}A_n(\hat{\theta}_{n,j} + \eta_j/\sqrt{n}) &= -\frac{1}{n}\int_0^1\int_0^1 v\ddot{\ell}_n(\hat{\theta}_{n,j} + uv\eta_j/\sqrt{n})dudv \\ &\approx -\frac{1}{n}\int_0^1\int_0^1 \sum_{i=1}^{n} E_{\theta_{0,j}}\dot{\Psi}(\theta_{0,j} \mid x_i, Y)vdudv \\ &\approx \frac{1}{2}\cdot\frac{1}{n}\sum_{i=1}^{n} I_i(\theta_{0,j} \mid x_i).\end{aligned}$$

Therefore, (102) becomes

$$\begin{aligned}\frac{L_n(\theta_j \mid \mathcal{D}_n)}{L_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)}w(\theta_j) &= \frac{L_n(\hat{\theta}_{n,j} + \eta_j/\sqrt{n} \mid \mathcal{D}_n)}{L_n(\hat{\theta}_{n,j} \mid \mathcal{D}_n)}w(\hat{\theta}_{n,j} + \eta_j/\sqrt{n}) \\ &\overset{a.s.}{\approx} \exp\left\{-\frac{1}{2}\eta_j^T\left(\frac{1}{n}\sum_{i=1}^{n} I_i(\theta_{0,j} \mid x_i)\right)\eta_j\right\}w(\theta_{0,j}).\end{aligned}$$

Hence, the posterior distribution of η_j is

$$\begin{aligned}p(\eta_j \mid \mathcal{D}_n) &\propto L_n(\hat{\theta}_{n,j} + \eta_j/\sqrt{n} \mid \mathcal{D}_n)w(\hat{\theta}_{n,j} + \eta_j/\sqrt{n}) \\ &\propto \exp\left\{-\frac{1}{2}\eta_j^T\left(\frac{1}{n}\sum_{i=1}^{n} I_i(\theta_{0,j} \mid x_i)\right)\eta_j\right\}.\end{aligned}$$

That means

$$p(\eta_j \mid \mathcal{D}_n) \overset{a.s.}{\approx} N\left(0, \left[\frac{1}{n}\sum_{i=1}^{n} I_i(\theta_{0,j} \mid x_i)\right]^{-1}\right).$$

Transforming from η_j back to θ_j gives Lemma 2. ■

A.2 INID Results – wrong model

Let $y_i \sim p_j(y \mid x_i, \theta_j)$ be INID and assume y has the true distribution P_T on Ω with the true density $p_T(y)$. Define the relative entropy distance between $p_T(y)$ and $1/n \sum_{i=1}^{n} p_j(y \mid x_i, \theta_j)$ as

$$KL\left[p_T(y) \| \frac{1}{n}\sum_{i=1}^{n} p_j(y \mid x_i, \theta_j)\right] = E\left[\log\left(\frac{p_T(y)}{\frac{1}{n}\sum_{i=1}^{n} p_j(y \mid x_i, \theta_j)}\right)\right] \tag{104}$$

where the expectations are taken w.r.t. the true distribution P_T.

Consider the quasi-log-likelihood

$$L_n(\theta_j \mid \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} p_j(y \mid x_i, \theta_j)$$

and let $\hat{\theta}_{n,j}$ be the quasi-MLE, i.e

$$\hat{\theta}_{n,j} = \arg\max_{\theta_j \in \Theta_j} L_n(\theta_j \mid \mathcal{D}_n). \tag{105}$$

Define

$$A(\theta_j \mid x^n) = E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_\ell \partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right] \text{ and}$$
$$B(\theta_j \mid x^n) = E\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta_\ell} \log p_j(y \mid x_i, \theta_j) \cdot \frac{\partial}{\partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right], \ell, k = 1, \ldots, d.$$

Now we consider the following list of hypotheses.

(i) The densities $p_j(y \mid x_i, \theta_j)$ are measurable in y for all $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$, and continuous in θ_j for all $y \in \Omega$, $x_i \in \mathcal{X}$ (Θ_j, $\mathcal{X}$ are compact sets).

(ii) $E[\log p_T(y)]$ exists; $|\log p_j(y \mid x_i, \theta_j)| \leq m(y)$ for all $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$ where m is integrable w.r.t. P_T; and θ_j^*, the parameter minimizes (104), is unique.

(iii) $\frac{\partial}{\partial\theta_\ell} \log p_j(y \mid x_i, \theta_j)$, $\ell = 1, \ldots, d$, are measurable functions of y for all $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$, and continuously differentiable functions of θ_j for all $y \in \Omega$, $x_i \in \mathcal{X}$.

(iv) $\left|\frac{\partial^2}{\partial\theta_\ell \partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right|$ and $\left|\frac{\partial}{\partial\theta_\ell} \log p_j(y \mid x_i, \theta_j) \cdot \frac{\partial}{\partial\theta_k} \log p_j(y \mid x_i, \theta_j)\right|$, $\ell, k = 1, \ldots, d$, are dominated by functions integrable w.r.t. P_T for all $y \in \Omega$, $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$.

(v) θ_j^* is interior to Θ_j; $A(\theta_j^* \mid x^n)$ is nonsingular; and θ_j^* is a regular point of $A(\theta_j \mid x^n)$ i.e $A(\theta_j \mid x^n)$ has constant rank in some open neighborhood of θ_j^*.

We have the following.

Lemma 3 (White (1982) for INID case) *Let $y_i \sim p_j(y \mid x_i, \theta_j)$ be INID and assume y has the true distribution P_T on Ω with the true density $p_T(y)$. Assume hypotheses (i) - (v). If $\hat{\theta}_{n,j}$ is the quasi-MLE as defined in* (105)*, then*

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_j^*) \xrightarrow{L} N(0, C(\theta_j^* \mid x^n)),$$

where θ_j^* is the parameter which minimizes (104) *and*

$$C(\theta_j^* \mid x^n) = A(\theta_j^* \mid x^n)^{-1} B(\theta_j^* \mid x^n) A(\theta_j^* \mid x^n)^{-1}.$$

Proof All the steps in the proof of White (1982) Theorem 3.2 can now be directly applied under Assumptions (i) - (v). ■

Consider the second list of hypotheses as below.

(i) The densities $p_j(y \mid x_i, \theta_j)$ are measurable in y for all $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$, and continuous in θ_j for all $y \in \Omega$, $x_i \in \mathcal{X}$ (Θ_j, $\mathcal{X}$ are compact sets).

(ii) $p_j(y \mid x_i, \theta_j) > 0$ a.e. w.r.t. P_T for all $\theta_j \in \Theta_j$, $x_i \in \mathcal{X}$.

(iii) For every $\theta_j \in \Theta_j$, there is an open neighborhood B of θ_j such that

$$E \sup_{\theta_j \in B, x_i \in \mathcal{X}} \{|\log p_j(y \mid x_i, \theta_j)|\} < \infty.$$

(iv) There is a positive integer m such that for any real number r, there is a co-compact subset D of Θ_j (i.e $\Theta_j \setminus D$ is compact) such that

$$E \sup_{\theta_j \in D, x_i \in \mathcal{X}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \log p_j(y \mid x_i, \theta_j) \right\} \leq r.$$

We have the following.

Lemma 4 (Berk (1966) for INID case) *Let $y_i \sim p_j(y \mid x_i, \theta_j)$ be INID and assume y has the true distribution P_T on Ω with the true density $p_T(y)$. Assume hypotheses (i) - (iv) and that the prior $w_j(\theta_j) > 0$ is continuous for all $\theta_j \in \Theta_j$. Then, the posterior $w_j(\theta_j \mid \mathcal{D}_n)$ is almost surely carried on the set*

$$A_{0,j} = \left\{ \theta_j^* \mid \theta_j^* \in \arg\min_{\theta_j \in \Theta_j} KL \left[p_T(y) \| \frac{1}{n} \sum_{i=1}^{n} p_j(y \mid x_i, \theta_j) \right] \right\}.$$

Proof All the steps in the proof of the main theorem in Berk (1966) can now be directly applied under Assumptions (i) - (iv). ■

Note: we don't have asymptotic normality for the posterior as in Lemma 2, but we can still obtain Prop. 2 if θ_j^* is unique as assumed in White (1982).

Appendix B. Uniformity Results

The two lemmata here are uniformity results for the sequences of random variables appearing in Theorem 7. These results are not surprising but do not seem to be in the standard literature or follow directly from it. Our first Lemma here shows that the expressions (63) are uniformly $o_P(1)$.

Lemma 5 *Assume*

(i) For any $j = 1, \ldots, J$, and n,

$$E_j(Y_n^2) = \int\int y^2 p_j(y_n \mid x, \theta_j) w_j(\theta_j) d\theta_j dy < \infty,$$

(ii) For each $j = 1, \ldots, J$ and n, the conditional densities $p_j(y_n \mid x, \theta_j)$ are equicontinuous in x, for $x \in K_1$ for each y and $\theta_j \in K_{2j}$ where K_1 and the K_{2j}'s are compact sets, and,

(iii) There exists $c > 0$ so that for any $j = 1, \ldots, J$ and n,

$$E_j(Y(x) - \hat{f}_j(x))^2 < c < \infty.$$

Then the expressions (63), i.e.,

$$(Y_i - \hat{f}_{j,-i}(x_i))^2 - (Y_i - \hat{f}_j(x_i))^2 = (\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2Y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i))$$

for $i = 1, \ldots, n$ are uniformly $o_P(1)$.

Proof By the Cauchy-Schwarz inequality,

$$\begin{aligned} (\hat{f}_j(x) - \hat{f}_{j,-i}(x))^2 &= (E_j(Y_{n+1} \mid \mathcal{D}_n) - E_j(Y_{n+1} \mid \mathcal{D}_{n,-i}))^2 \\ &\leq 2\left(E_j^2(Y_{n+1} \mid \mathcal{D}_n) + E_j^2(Y_{n+1} \mid \mathcal{D}_{n,-i})\right) \\ &\leq 2\left(E_j(Y_{n+1}^2 \mid \mathcal{D}_n) + E_j(Y_{n+1}^2 \mid \mathcal{D}_{n,-i})\right). \end{aligned} \tag{106}$$

So, by Assumption (i) and Billingsley (2012) (Lemma p. 498), we have that the right hand side of (106) is a uniformly integrable sequence since the sequence of x_i's in K_1 is regarded as a countable collection of fixed design points. Assumption (ii) together with Prop. 2 gives

$$E_j(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))^2 \to 0 \text{ as } n \to \infty \tag{107}$$

uniformly in the x_i's.

For $\epsilon > 0$, Markov's inequality gives

$$\begin{aligned} &P_j[(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2Y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i)) > \epsilon] \\ &\leq \frac{1}{\epsilon} E_j[(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2Y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i))] \\ &\leq \frac{1}{\epsilon}\left[E_j(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))^2 E_j(2Y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i))^2\right]^{1/2} \\ &\leq \frac{1}{\epsilon}\left\{E_j(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))^2 \left[2E_j(Y_i - \hat{f}_{j,-i}(x_i))^2 + 2E_j(Y_i - \hat{f}_j(x_i))^2\right]\right\}^{1/2} \end{aligned} \tag{108}$$

Therefore, from Assumption (iii), (107), and (108), for each $\eta > 0, \epsilon > 0$ there exists $N(\eta, \epsilon)$ such that

$$\sup_{x_i} P_j[(\hat{f}_j(x_i) - \hat{f}_{j,-i}(x_i))(2Y_i - \hat{f}_{j,-i}(x_i) - \hat{f}_j(x_i)) > \epsilon] < \eta, \text{ for all } n > N(\eta, \epsilon),$$

i.e., the expressions (63) are uniformly $o_P(1)$. ■

Second, the following lemma shows that a sum of n uniformly $o_P(1)$ random variables is $o_P(n)$; this may be of more general interest.

Lemma 6 *If the sequence of random variables $\{Z_{i,n}, i = 1, \ldots, n\}$ is uniformly $o_P(1)$ then the sum $T_n = \sum_{i=1}^n Z_{i,n} = o_P(n)$.*

Proof Since $Z_{i,n} = o_P(1)$ uniformly, for each $\eta > 0, \epsilon > 0$ there exists $N(\eta, \epsilon)$ such that for any $\xi > 0$,

$$\sup_i P(|Z_{i,n}| > \epsilon - \frac{\xi}{n}) < \frac{\eta}{n}, \text{ for all } n > N(\eta, \epsilon). \tag{109}$$

Consider

$$\begin{aligned} P(|T_n| > n\epsilon) &= P\left(\left|\frac{\sum_{i=1}^n Z_{i,n}}{n}\right| > \epsilon\right) \\ &\leq P\left(\sum_{i=1}^n \frac{|Z_{i,n}|}{n} > \epsilon\right) \\ &\leq P\left(\sum_{i=1}^{N(\eta,\epsilon)} |Z_{i,n}| + \sum_{i=N(\eta,\epsilon)+1}^{n} |Z_{i,n}| > n\epsilon\right). \end{aligned} \tag{110}$$

Since $\sum_{i=1}^{N(\eta,\epsilon)} |Z_{i,n}| \xrightarrow{P} 0$, for any $\xi > 0$, the right hand side of (110) is bounded by

$$\begin{aligned} P\left(\xi + \sum_{i=N(\eta,\epsilon)+1}^{n} |Z_{i,n}| > n\epsilon\right) &= P\left[\sum_{i=N(\eta,\epsilon)+1}^{n} |Z_{i,n}| > \left(\epsilon - \frac{\xi}{n}\right) n\right] \\ &\leq P\left[\cup_{i=N(\eta,\epsilon)+1}^{n} \left(|Z_{i,n}| > \frac{\left(\epsilon - \frac{\xi}{n}\right) n}{n - N(\eta, \epsilon)}\right)\right] \\ &\leq \sum_{i=N(\eta,\epsilon)+1}^{n} P\left[|Z_{i,n}| > \frac{\left(\epsilon - \frac{\xi}{n}\right) n}{n - N(\eta, \epsilon)}\right]. \end{aligned}$$

Using $(\epsilon - \xi/n)/(1 - N(\eta, \epsilon)/n) > \epsilon - \xi/n$ gives the new bound

$$\sum_{i=N(\eta,\epsilon)+1}^{n} P\left(|Z_{i,n}| > \epsilon - \frac{\xi}{n}\right) \leq n \sup_i P\left(|Z_{i,n}| > \epsilon - \frac{\xi}{n}\right).$$

Therefore, from (109), for each $\eta > 0, \epsilon > 0$ there exists $N(\eta, \epsilon)$ such that

$$P(|T_n| > n\epsilon) < \eta, \text{ for all } n > N(\eta, \epsilon),$$

i.e., $T_n = \sum_{i=1}^{n} Z_{i,n} = o_P(n)$. ■

References

P. Bellec and A. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *J. Machince Learning Res.*, 16:1879–1893, 2015.

R. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, 37:51–58, 1966.

J. Bernardo and A. Smith. *Bayesian Theory.* John Wiley & Sons, Chichester, 2000.

Patrick Billingsley. *Probability and Measure.* Wiley, New Jersey, 2012.

Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996a.

Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996b.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Peter Buhlmann and Bin Yu. Analyzing bagging. *Ann. Statist.*, 30:927–961, 2002.

F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for regression learning, 2004. URL `https://arxiv.org/pdf/math/0410214.pdf`.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, Cambridge, 2006.

H. Chipman, E. George, and McCulloch. The practical implementation of bayesian model selection. *Model selection, Institute of mathematical statistics lecture notes*, pages 65–116, 2001.

B. Clarke. *Information-Theoretic Asymptotics of Bayes Methods.* PhD thesis, Department of Statistics, University of Illinois, 1989.

B. Clarke. Asymptotic normality of the posterior in relative entropy. *IEEE Trans. Inform Theory*, pages 165–176, 1999.

B. Clarke. Bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research*, pages 683–712, 2003.

B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. Technical report, 1988.

B. Clarke and J. Clarke. *Predictive Statistics: Analysis and Inference Beyond Models.* Cambridge University Press, Cambridge, 2018.

B. Clarke and E. Fokoue. Bias-variance trade-off for prequential model list selection. *Stat. Papers*, 52:813–833, 2011.

Robert Clemen. Combining forecasts: A review and annotated bibliography. *Int'l J. Forecasting*, 5:559–583, 1989.

M. Clyde and E. Iversen. Bayesian model averaging in the M-open framework. In P. Damien, P. Dellaportas, N. Polson, and D. Stephens, editors, *Bayesian Theory and Applications*, pages 484–498, Oxford, 2013. Oxford University Press.

Melise Clyde and Edward George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

A. Dalalyan and J. Salmon. Sharp orcale inequalities for aggregation of affince estimators. *Ann. Stat.*, pages 2327–2355, 2012.

Phillip Dawid. The prequential approach. *Journal of the Royal Statistical Society*, 147: 287–292, 1984.

Phillip Dawid and Vladimir Vovk. Prequential probability: principles and properties. *Bernoulli*, 5:125–162, 1999.

N. DeBruijn. *Asymptotic Methods in Analysis.* Dover, NY, 1958.

David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Association*, 57:45–97, 1995.

T. Ferguson. *A Course in Large Sample Theory.* Chapman & Hall/CRC, Boca Raton, FL, 33431, 1996.

Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Computer and System Sciences*, 55:119–139, 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A statistical view of boosting. *The Annals of Statistics*, 28:337–407, 2000.

Francis Galton. The wisdom of crowds. *Nature*, 75:450–451, 1907.

S. Geisser. *Predictive inference.* Chapman and Hall, London, 1993.

E. George. Sampling considerations for model averaging and model search. Invited discussion of 'model averaging and model search'. *Bayesian Statistics*, 6:175–177, 1999.

Peter Hall. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York, 1992.

B. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

J. Hartigan. Asymptotic normality of posterior distributions. In *Bayes Theory*, pages 107–118, NY, 1983. Springer.

B. Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.*, 42:1977–1991, 1971.

T. Hu, A. Rosalsky, and A. Volodin. On convergence properties of sums of dependent random variables under second moment and covariance restrictions. *Stat. Prob. Letters*, pages 1999–2005, 2008.

S. Kong and B. Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Stat. Sin.*, 24:25–42, 2014.

T. Le and B. Clarke. Using the Bayesian shtarkov solution for predictions. *Computational Statistics and Data Analysis, http://dx.doi.org/10.1016/j.csda.2016.06.018*, 104:183–196, 2016.

T. Le and B. Clarke. A Bayes interpretation of stacking in M-complete and M-open settings. *Bayesian Analysis*, 12:807–829, 2017.

T. Le and B. Clarke. On the interpretation of ensemble classifiers in terms of Bayes classifiers. *J. Classif.*, 35:198–229, 2018.

E. Leamer. *Specification searches: ad hoc inference with nonexperimental data.* Wiley, New York, 1978.

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13:1000–1022, 2007.

G. Lecué and G. Mitchel. Oracle inequalities for cross-validation type procedures, 2010. URL `http://www.cmapx.polytechnique.fr/~lecue/LecMit-May2010.pdf`.

J. Lederer, L. Yu, and I. Gaynanova. Oracle inequalities for high-dimensional prediction. *Bernoulli*, 29:1225-1255, 2019.

David Madigan and Adrian Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

G. Maillard, S. Arlot, and M. Lerasle. Aggregated hold-out. *JMLR*, 21:1–55, 2021.

L. Montuelle and E. Pennec. Pac-bayesian aggregation of affine estimators. 2018. URL `https://hal.inria.fr/hal-01070805v2`.

W. Newey and D. McFadden. Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics Vol. IV*, pages 2111–2245. Elsevier Science, BV, 2012.

Luca Onorante and Adrian E. Raftery. Dynamic model averaging in large model spaces using dynamic occam's window. *arXiv:1410.7799*, 2014.

M. Ozay and F. T. Yarman Vural. A new fuzzy stacked generalization technique and analysis of its performance. *arXiv:1204.0171*, 2012.

Adrian Raftery and Yingye Zheng. Discussion on performance of bayesian model averaging. *J. Amer. Stat. Assoc.*, 98:931–938, 2003.

Adrian Raftery, David Madigan, and C. Volinsky. *Accounting for model uncertainty in survival analysis improves predictive performance.* Bayesian statistics 5, Oxford University Press, 1996.

Suhansini Subba Rao. https://www.stat.tamu.edu/ suhasini/teaching613/bootstrap.pdf, 2017.

N. Reid, R. Mukerjee, and D.A.S. Fraser. Some aspects of amtching priors. In *Mathematical Statistics and Applications: Festschrift for Constance van Eeden*, volume 42, pages 31–43. IMS Lecture Notes-Monograph Series, 2003.

G. Ridgeway, D. Madigan, and T. Richardson. Boosting methodology for regression problems. In D. Heckerman and J. Whittaker, editors, *Proceedings of Artificial Intelligence and Statistics '99*, page 152–161, 1999a.

Greg Ridgeway, David Madigan, and Thomas Richardson. Boosting methodology for regression problems. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.207*, 1999b.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, 30:629–636, 1984.

L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39, 2010.

R. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

R. Schapire and Y. Freund. *Boosting: Foundations and Algorithms.* MIT Press, MA, 2012.

G. Shmueli. To explain or predict? *Stat. Sci.*, 25:289–310, 2010.

J. Sill, G. Takacs, L. Mackey, and D. Lin. Feature-weighted linear stacking. *arXiv:0911.0460*, 2009.

Kostas Skouras and Phillip Dawid. On efficient point prediction systems. *Journal of the Royal Statistical Society*, 60:765–780, 1998.

Kostas Skouras and Phillip Dawid. On efficient probability forecasting systems. *Biometrika*, 86:765–784, 1999.

P. Smyth and D. Wolpert. Linearly combining density estimators via stacking. *Machine Learning Journal*, 36:59–83, 1999.

Carolin Strobl, James Malley, and Gerhard Tutz. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14:323–348, 2009.

Kai Ming Ting and Ian Witten. Issues in stacked generalization. *Journal of Artificial Intelligent Research*, 10:271–289, 1999.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 2020.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

Y. Yang and D. Pati. Bayesian model selection consistency and oracle inequality with intractable marginal likelihood, 2017. URL `https://arxiv.org/abs/1701.00311`.

Qingzhao Yu, Steven N. MacEachern, and Mario Peruggia. Clustered bayesian model averaging. *Bayesian Analysis*, 8:883–908, 2013.